

LONGMANS' PSYCHOLOGY SERIES

GENERAL EDITOR

JOHN A. MCGEOCH

*Professor of Psychology
Wesleyan University, Connecticut*

CHILD PSYCHOLOGY

By MARGARET WOOSTER CURTI

Research Associate in Educational Psychology, Teachers College

GENERAL EXPERIMENTAL PSYCHOLOGY

By ARTHUR GILBERT BILLS

Head of Psychology Department, University of Cincinnati

STATISTICS IN PSYCHOLOGY AND EDUCATION

By HENRY E. GARRETT

Associate Professor of Psychology, Columbia University

PSYCHOLOGY—A STUDY OF MENTAL ACTIVITY

By HARVEY A. CARR

AN INTRODUCTION TO SPACE PERCEPTION

By HARVEY A. CARR

MENTAL HYGIENE AND EDUCATION

By MANDEL SHERMAN

*Associate Professor of Educational Psychology, and Psychiatrist,
The Orthogenic School, University of Chicago*

MENTAL CONFLICTS AND PERSONALITY

By MANDEL SHERMAN

THE PROBLEM OF STUTTERING

A DIAGNOSIS AND A PLAN OF TREATMENT

By JOHN MADISON FLETCHER

Professor of Psychology, Tulane University

THE MENTAL LIFE

By CHRISTIAN A. RUCKMICK

**STATISTICS IN PSYCHOLOGY
AND EDUCATION**

“If we take in our hand any volume . . . let us ask, *Does it contain any abstract reasoning concerning quantity or number?* No. *Does it contain any experimental reasoning concerning matter of fact and existence?* No. Commit it then to the flames: for it can contain nothing but sophistry and illusion!”

Hume, David, *An Enquiry Concerning Human Understanding*, (1777).

STATISTICS IN PSYCHOLOGY AND EDUCATION

BY

HENRY E. GARRETT, PH.D.

ASSOCIATE PROFESSOR OF PSYCHOLOGY, COLUMBIA UNIVERSITY

WITH AN INTRODUCTION BY

R. S. WOODWORTH

PROFESSOR OF PSYCHOLOGY
COLUMBIA UNIVERSITY

SECOND EDITION

LONGMANS, GREEN AND CO.

NEW YORK • LONDON • TORONTO

1941

LONGMANS, GREEN AND CO.
55 FIFTH AVENUE, NEW YORK
221 EAST 20TH STREET, CHICAGO

LONGMANS, GREEN AND CO. LTD.
43, ALBERT DRIVE, LONDON, S.W. 19
17 CHITTARANJAN AVENUE, CALCUTTA
NICOL ROAD, BOMBAY
36A MOUNT ROAD, MADRAS

LONGMANS, GREEN AND CO.
215 VICTORIA STREET, TORONTO

GARRETT
STATISTICS IN PSYCHOLOGY AND EDUCATION

COPYRIGHT · 1926 · 1937
BY LONGMANS, GREEN AND CO.
ALL RIGHTS RESERVED, INCLUDING THE
RIGHT TO REPRODUCE THIS BOOK, OR
ANY PORTION THEREOF, IN ANY FORM

First Edition January 1926
Reprinted November 1926
October 1927, June 1929, July 1930
August 1931, August 1932, August 1933
November 1934, July 1935
Second Edition rewritten June 1937
Reprinted September 1937
June 1938, March 1939
July 1940, August 1941

PRINTED IN THE UNITED STATES OF AMERICA

INTRODUCTION

MODERN problems and needs are forcing statistical methods and statistical ideas more and more to the fore. There are so many things we wish to know which cannot be discovered by a single observation, or by a single measurement. We wish to envisage the behavior of a man who, like all men, is rather a variable quantity, and must be observed repeatedly and not once for all. We wish to study the social group, composed of individuals differing one from another. We should like to be able to compare one group with another, one race with another, as well as one individual with another individual, or the individual with the norm for his age, race or class. We wish to trace the curve which pictures the growth of a child, or of a population. We wish to disentangle the interwoven factors of heredity and environment which influence the development of the individual, and to measure the similarly interwoven effects of laws, social customs and economic conditions upon public health, safety and welfare generally. Even if our statistical appetite is far from keen, we all of us should like to know enough to understand, or to withstand, the statistics that are constantly being thrown at us in print or conversation — much of it pretty bad statistics. The only cure for bad statistics is apparently more and better statistics. All in all, it certainly appears that the rudiments of sound statistical sense are coming to be an essential of a liberal education.

Now there are different orders of statisticians. There is, first in order, the mathematician who invents the method for performing a certain type of statistical job. His interest, as a mathematician, is not in the educational, social or psychological problems just alluded to, but in the problem of devising instruments for handling such matters. He is the tool-maker of the

statistical industry, and one good tool-maker can supply many skilled workers. The latter are quite another order of statisticians. Supply them with the mathematician's formulas, map out the procedure for them to follow, provide working charts, tables and calculating machines, and they will compute from your data the necessary averages, probable errors and correlation coefficients. Their interest, as computers, lies in the quick and accurate handling of the tools of the trade. But there is a statistician of yet another order, in between the other two. His primary interest is psychological, perhaps, or it may be educational. It is he who has selected the scientific or practical problem, who has organized his attack upon the problem in such fashion that the data obtained can be handled in some sound statistical way. He selects the statistical tools to be employed, and, when the computers have done their work, he scrutinizes the results for their bearing upon the scientific or practical problem with which he started. Such an one, in short, must have a discriminating knowledge of the kit of tools which the mathematician has handed him, as well as some skill in their actual use.

The reader of the present book will quickly discern that it is intended primarily for statisticians of the last-mentioned type. It lays out before him the tools of the trade; it explains very fully and carefully the manner of handling each tool; it affords practice in the use of each. While it has little to say of the tool-maker's art, it takes great pains to make clear the use and limitations of each tool. As any one can readily see who has tried to teach statistics to the class of students who most need to know the subject, this book is the product of a genuine teacher's experience, and is exceptionally well adapted to the student's use. To an unusual degree, it succeeds in meeting the student upon his own ground.

R. S. WOODWORTH

COLUMBIA UNIVERSITY
(1926)

PREFACE

TO SECOND EDITION

IN THE ten years which have elapsed since the publication of the first edition of this book, there have developed, I think, two fairly distinct trends in the application of statistical method to problems in psychology and education. In the first place, there has been a tremendous growth in the use of statistical methods; and this has been true in spite of the inclination shown by some psychologists to substitute, or to advocate the substitution of, qualitative for quantitative methods. In the second place, the application of quantitative methods in psychology and education has, I think, become more discriminating and more judicious. The student of today who plans to carry on research must know more and better statistics than the student of ten years ago, and, fortunately, many of them do.

In rewriting this book I have tried to take account of both developments mentioned. I have added a number of useful techniques not included in the old book and have tried to indicate more precisely where a given method does — and does not — apply. The treatment of percentiles, of comparable scores, of reliability and validity of tests has been expanded; and new material has been added to the chapters dealing with the normal probability curve, sampling and reliability of measures, and correlational methods. Breaking down the chapters of the old book into smaller and more comprehensible units should improve the teachability of the new book.

I have profited greatly from the criticisms of those who have used the book in the classroom and have followed their suggestions wherever I could conscientiously do so. Thus, owing to many requests, and for other reasons, I have adopted the mathematical definition of a score throughout the new book.

I still feel that much can be said in favor of the "psychological" notion of a score which I used before. But such usage is anathema, apparently, to the mathematically trained teacher, and a long discussion of the merits and demerits of this or that method probably serves to confuse the student rather than to enlighten him.

For advice and help in many ways I am especially indebted to Professor J. W. Dunlap of Fordham University; to Professor J. F. Walker of the University of Arizona; to Dr Anne Anastasi of Barnard College; and to Mr. W. G. Madow of Columbia University. I am grateful also to Ruth E. Perl, Irene Gansl, Mary T. Wilson and Mildred B. Garrett for computations and for aid in the preparation of the manuscript.*

HENRY E. GARRETT

Department of Psychology
Columbia University

*Table 52 has been copied from R. A. Fisher's "Statistical Methods for Research Workers," 1936, Oliver and Boyd, Edinburgh and London, with kind permission of the publishers.

CONTENTS

CHAPTER I

THE FREQUENCY DISTRIBUTION

SECTION	PAGE
I. MEASURES IN GENERAL	1
II. DRAWING UP A FREQUENCY DISTRIBUTION . . .	4
III. STANDARDS OF ACCURACY IN COMPUTATION . . .	10

CHAPTER II

MEASURES OF CENTRAL TENDENCY

✓I. CALCULATION OF MEASURES OF CENTRAL TENDENCY	17
II. CALCULATION OF THE MEAN BY THE "ASSUMED MEAN" OR SHORT METHOD	26
III. WHEN TO USE THE VARIOUS MEASURES OF CENTRAL TENDENCY	29

CHAPTER III

MEASURES OF VARIABILITY

✓I. CALCULATION OF MEASURES OF VARIABILITY . .	34
II. CALCULATION OF THE <i>AD</i> AND <i>SD</i> BY THE SHORT METHOD	44
III. THE COEFFICIENT OF VARIATION, <i>V</i>	51
IV. THE SHORT METHOD APPLIED TO DISCRETE SERIES	55
✓V. WHEN TO USE THE VARIOUS MEASURES OF VARIABILITY	59

CHAPTER IV

GRAPHIC METHODS AND PERCENTILES

✓I. THE GRAPHIC REPRESENTATION OF THE FREQUENCY DISTRIBUTION	62
II. OTHER GRAPHICAL METHODS	89

CHAPTER V

✓ *THE NORMAL PROBABILITY CURVE*

SECTION	PAGE
I. THE MEANING AND IMPORTANCE OF THE NORMAL DISTRIBUTION	98
II. TABLES OF FREQUENCIES OF THE NORMAL PROBABILITY DISTRIBUTION	108
III. THE MEASUREMENT OF DIVERGENCE FROM NORMALITY.	115
IV. WHY OBTAINED DISTRIBUTIONS OFTEN DEVIATE FROM THE NORMAL FORM	128

CHAPTER VI

*APPLICATIONS OF THE NORMAL
PROBABILITY CURVE*

I. PROBLEMS INVOLVING PERCENTAGES OF AREA WITHIN DIFFERENT PARTS OF THE NORMAL DISTRIBUTION	132
II. THE SCALING OF TEST ITEMS	143
III. THE TRANSMUTATION OF MEASURES BY RELATIVE POSITION INTO UNITS OF AMOUNT	157

CHAPTER VII

*COMPARABLE MEASURES; COMBINING TEST
SCORES AND DISTRIBUTIONS*

I. METHODS OF RENDERING TEST SCORES COMPARABLE	178
II. THE MEAN AND SIGMA FROM COMBINED DISTRIBUTIONS	191

CHAPTER VIII

SAMPLING AND RELIABILITY

I. THE MEANING OF RELIABILITY	198
II. THE RELIABILITY OF MEASURES OF CENTRAL TENDENCY	200

CONTENTS

xi

SECTION	PAGE
III. THE RELIABILITY OF MEASURES OF VARIABILITY	208
IV. THE RELIABILITY OF THE DIFFERENCE BETWEEN TWO MEASURES	210
V. THE RELIABILITY OF CERTAIN OTHER MEASURES	226
VI. VARIOUS PROBLEMS WHICH INVOLVE MEASURES OF RELIABILITY	231
VII. SAMPLING AND RELIABILITY	242

CHAPTER IX

LINEAR CORRELATION

I. WHAT IS MEANT BY CORRELATION	251
II. THE COEFFICIENT OF CORRELATION	255
III. THE CALCULATION OF THE COEFFICIENT OF CORRE- LATION BY THE PRODUCT-MOMENT METHOD	265
IV. THE RELIABILITY OF THE COEFFICIENT OF CORRE- LATION	280

CHAPTER X

REGRESSION AND PREDICTION

I. THE REGRESSION EQUATIONS	289
II. THE RELIABILITY OF PREDICTIONS MADE FROM REGRESSION EQUATIONS	300
III. THE EFFECT UPON THE CORRELATION COEFFICIENT OF THE RANGE OF TALENT IN THE GROUP	303
IV. THE COMPLETE SOLUTION OF A CORRELATION PROBLEM	305

CHAPTER XI

THE RELIABILITY AND VALIDITY OF TEST SCORES

I. RELIABILITY OF TEST SCORES	311
II. VALIDITY OF TEST SCORES	324
III. THE ESTIMATION OF TRUE MEASURES	331

CHAPTER XII

*THE INTERPRETATION OF THE COEFFICIENT
OF CORRELATION*

SECTION	PAGE
I. VARIOUS INTERPRETATIONS OF THE COEFFICIENT OF CORRELATION	342

CHAPTER XIII

FURTHER METHODS OF CORRELATION

I. METHODS OF MEASURING CORRELATION WHICH TAKE ACCOUNT ONLY OF RELATIVE POSITION OR RANK	359
II. METHODS OF MEASURING CORRELATION OR ASSO- CIATION WHEN THE DATA ARE GROUPED INTO CLASSES OR CATEGORIES	366
III. CURVILINEAR OR NON-LINEAR RELATIONSHIP . .	393

CHAPTER XIV

PARTIAL AND MULTIPLE CORRELATION

I. THE MEANING OF PARTIAL AND MULTIPLE COR- RELATION	409
II. AN ILLUSTRATIVE CORRELATION PROBLEM INVOLV- ING THREE VARIABLES	412
III. GENERAL FORMULAS FOR USE IN PARTIAL AND MULTIPLE CORRELATION	420
IV. OUTLINE OF FORMULAS NEEDED IN CORRELATION PROBLEMS INVOLVING (a) FOUR VARIABLES, AND (b) FIVE VARIABLES	433
V. A MULTIPLE CORRELATION PROBLEM INVOLVING FOUR VARIABLES	437
VI. THE VALUE AND USE OF PARTIAL AND MULTIPLE CORRELATION	445
VII. SPURIOUS CORRELATION	457

CONTENTS

xiii

	PAGE
REFERENCES	465
REFERENCE TABLES	467
TABLE OF SQUARES AND SQUARE ROOTS OF THE NUMBERS FROM 1-1000	476
INDEX	487

TO THE INSTRUCTOR

This book contains more material than can, perhaps, be covered thoroughly in a one semester course. The following selection of topics is suggested, therefore, as meeting the requirements of a course in "minimum essentials."

Chapters I, II, and III

Chapter IV (I)

Chapter V (I and II)

Chapter VI (II)

Chapter VII

Chapter VIII (I, II, III, IV, and VII)

Chapter IX

Chapter X (I and II)

Chapter XI (I)

Chapter XII (I, 1)

Chapter XIII (I)

Chapter XIV (I, II, and VII)

STATISTICS IN PSYCHOLOGY
AND EDUCATION

CHAPTER I

THE FREQUENCY DISTRIBUTION

I. MEASURES IN GENERAL

1. Continuous and Discrete Series

IN the measurement of mental and social traits, most of the facts with which we deal fall into what are known as *continuous series*. A continuous series is one which is theoretically capable of any degree of subdivision, although in practice divisions smaller than some convenient unit are rarely employed. Measurements of general intelligence illustrate scores which fall into continuous series. I.Q.'s, for example, may be thought of as increasing by increments of 1 on a scale which extends from the idiot to the genius. But there is no reason, at least theoretically, why with more refined methods of measurement we should not be able to get I.Q.'s of 100.8 or even of 100.83. Physical measures such as height, weight, and cephalic index, as well as nearly all capacities measured by mental and educational tests, fall into continuous series. Within the range of the scale employed, any measure, integral or fractional, may exist and have meaning. When gaps occur in a truly continuous series, these are to be attributed to a failure to measure enough cases, to the relative crudity of our measuring instrument, or to some other reason of the same sort, rather than to the fact that no measures exist within the gaps.

There are, however, quantities which do not fall into continuous series. A salary scale in a department store, for instance, may run from \$10 per week to \$20 per week in units of \$1; no one receives, let us say, \$17.53 per week. Or, to take another example, the average family in a certain locality may work out mathematically to have 4.57 children, although there

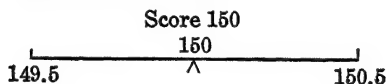
is obviously a real gap between four children and five children. Series like these, which exhibit real gaps, are called *discrete* or *discontinuous*.

It is fortunate, at least from the standpoint of the beginning student of statistical methods, that nearly all of the measurements with which we deal in psychology and education are continuous or may be profitably treated as continuous. This makes it possible for us to concern ourselves for the present, at least, almost entirely with methods of handling continuous data, postponing the discussion of discrete data to a later page.

In the following section we shall define more precisely just what is meant by a *score* in a continuous series, and then show how scores may be classified into what is called a *frequency distribution*.

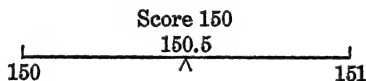
2. The Meaning of a Score in a Continuous Series .

Scores or other measures in a continuous series are to be thought of as *distances* along a scale, quantities between two limits, rather than as single points on a scale. An inch is a magnitude between two divisions on a foot-rule; and, in like manner, a score in a mental test is a unit distance between two divisions upon some defined mental scale. Let us consider, for example, a score of 150 upon the Army Alpha Intelligence Examination. This value represents the interval 149.5 to 150.5 on a scale of general intelligence. The exact midpoint of the score interval is 150 as shown here.



Other scores are to be interpreted in the same way as intervals upon some defined scale. A score of 8 on the Thorndike Handwriting Scale, for instance, includes all values from 7.5 to 8.5; i e., any value from a point .5 unit *below* 8, to .5 unit *above* 8. Hence, 7.7, 8.0, and 8.4 may all be scored 8. An interval extending from .5 unit below to .5 unit above the given value is the ordinary mathematical meaning of a single score.

There is another and somewhat different meaning which a test score may have. According to this second view, a score of 150 on Army Alpha represents any value on the scale *between* 150 and 151. A score of 150, in other words, implies that an individual has completed *at least* 150 items correctly, but not 151. Any fractional value greater than 150, but less than 151, e.g., 150.3 or 150.8, since it falls within the interval 150–151 may be scored simply as 150. The middle of the score interval is 150.5.



Both of these ways of defining a score are useful. Which to use will depend upon the way in which the test is scored and on the meaning of the units of measurement employed. If each of ten boys is recorded as having a height of 64 inches this will ordinarily mean that their heights fall between 63.5 and 64.5 inches (middle value 64 in.), and not between 64 and 65 inches (middle value 64.5 in.). On the other hand, the ages of twenty-five children, all recorded as being 9 years old, will most probably lie between 9 and 10 years; will be greater than 9 and less than 10 years (middle value 9.5). But "9 years old" must be taken in many studies to mean 8.5 to 9.5 years with a middle value of 9 years. The main point to remember is that results obtained from treating scores under our second definition will always be .5 unit higher than results obtained when scores are taken under the first or mathematical definition. The student will often have to decide, perhaps somewhat arbitrarily, which meaning a score should have. As a general rule it is safer to take the mathematical meaning of a score unless clearly indicated otherwise. This will be the method followed throughout this book. That is, scores of 62 and 231, say, will usually mean 61.5 up to 62.5, and 230.5 up to 231.5, and not 62 up to 63, and 231 up to 232.

II. DRAWING UP A FREQUENCY DISTRIBUTION

1. The Classification of Measures

Data collected from tests and experiments are often a series of numbers with little meaning or significance until they have been rearranged or classified in a systematic way. The first task that confronts us, then, is the organization of our material and this leads naturally to a grouping of the measures or scores into classes or categories. The procedure in grouping falls under three main heads:

(1) Determination of the *range* or the interval between the largest and smallest measures. The range is found by subtracting the smallest from the largest measure.

(2) Decision as to the *number* and *size* of the groups to be used in classification. The number and size of these *step-* or *class-intervals* will depend upon the range of scores and the kind of measures with which we are dealing.

(3) Tabulation of the separate measures within their proper *step-* or *class-intervals*.

These three principles of classification are illustrated in Table 1. The figures in this table represent the Army Alpha scores earned by 50 college men. Since the highest score is 197, and the lowest 142, the range (197-142) is exactly 55 points. In deciding upon the number of steps or classes to be used in grouping a good general rule is to select by trial an interval which will yield not more than 20 nor less than 10 classes.*

The number of steps or classes which a given range will yield can be determined approximately (within one step) by dividing the range by the step-interval tentatively chosen. In the present problem, 55 (the range) divided by 5 (the step-interval) gives 11, which is one less than the actual number of steps, namely, 12. A step-interval of 3 units will yield 19 steps; a step-interval of 10 units, 6 steps.

The tabulation of the separate scores within their *step-* or *class-intervals* is shown in Table 1. In the first column of this

*This rule must often be broken when the number of scores is small.

TABLE 1
THE TABULATION OF FIFTY ARMY ALPHA SCORES MADE BY
COLLEGE STUDENTS

1. The original scores ungrouped

185	166	176	145	166	191	177	164	171	174
147	178	176	# 142	170	158	171	167	180	178
173	148	168	187	181	172	165	169	173	184
175	156	158	187	156	172	162	193	173	183
* 197	181	151	161	153	172	162	179	188	179

* Highest score

Lowest score

2. The same fifty scores grouped into a frequency distribution

(1) Step- or Class- Intervals	(2) Tallies	(3) <i>f</i> (frequency)
195 up to 200	/	1
190 " " 195	//	2
185 " " 190	///	4
180 " " 185	////	5
175 " " 180	/////	8
170 " " 175	////	10
165 " " 170	///	6
160 " " 165	////	4
155 " " 160	////	4
150 " " 155	///	2
145 " " 150	//	3
140 " " 145	/	1
		$N = 50$

table, the class-intervals have been listed serially from the smallest score at the bottom of the column to the largest score at the top. Each class-interval comprises exactly five scores. The first interval "140 up to 145" begins with score 140 and ends with 144, thus including the five scores 140, 141, 142, 143, and 144. The second interval "145 up to 150" begins with 145 and ends with 149, i.e., at score 150 on the scale. The last interval "195 up to 200" begins with score 195 on the scale and ends at score 200, thus including the scores 195, 196, 197, 198, 199. In column 2, marked "Tallies," the separate scores have been listed opposite their proper intervals. The first score, 185, is represented by a tally placed opposite step-interval "185 up to 190"; the second score, 147, by a tally placed opposite step-

interval "145 up to 150"; and the third score, 173, by a tally placed opposite "170 up to 175." The remaining scores have been tabulated in the same way. When all fifty scores have been listed, the total number of tallies on each step-interval (i.e., the frequency) is written in column 3, headed f (frequency). The sum of the f column is called N . When the total frequency within each step-interval has been tabulated opposite the proper step, as shown in column 3, our fifty Army Alpha scores are arranged in a *frequency distribution*.

The student will note that the lower limit of the first step in the distribution (140 up to 145) has been set at 140 although the lowest score in the series is 142. When the interval selected for tabulation is five units it facilitates tabulation as well as the computations which come later, if the lower limit of the first step-interval, and, accordingly, of each successive step-interval, is a multiple of five. A step-interval "142 up to 147" is just as good theoretically as a step-interval "140 up to 145"; but the second is easier to handle from the standpoint of the arithmetic involved.

2. Three Methods of Describing the Limits of the Step-Intervals in a Frequency Distribution

Table 2 illustrates three ways of expressing the limits of the step-intervals in a frequency distribution. In (A), the interval "140 up to 145" means, as we have already learned, that all scores from 140 up to but not including 145 fall within this grouping. The step-intervals in (B) cover the same distances as in (A), but the upper and lower limits of each step are defined more exactly. We have found (p. 2) that a score of 140 in a continuous series ordinarily means the interval 139.5-140.5; and that a score of 144 means 143.5-144.5. Accordingly, to express precisely the fact that a step *begins* with 140 and *ends* with 144, we may write 139.5 (the beginning of score 140) as the lower limit, and 144.5 (end of score 144) as the upper limit of this step. The step-intervals in (C) express the same facts more clearly than in (A) and less exactly than in (B). Thus,

"140-144" means that this step begins *with* score 140 and ends *with* score 144; but the precise limits of the step are not given. The diagram below will show how (A), (B), and (C) are three ways of expressing identically the same facts:

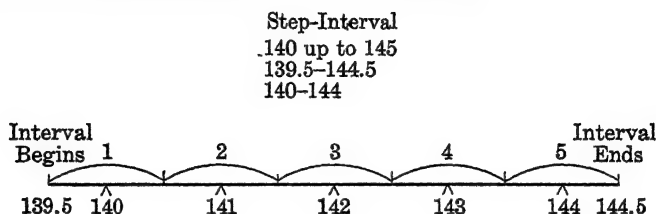


TABLE 2

THREE METHODS OF GROUPING SCORES INTO A
FREQUENCY DISTRIBUTION

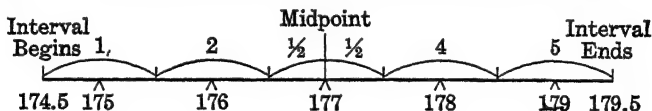
(The data are the 50 Army Alpha scores tabulated in Table 1, p. 5)

(A)			(B)			(C)		
Step-Intervals	Mid-point	f	Step-Intervals	Mid-point	f	Step-Intervals	Mid-point	f
195 up to 200	197	1	194.5-199.5	197	1	195-199	197	1
190 " " 195	192	2	189.5-194.5	192	2	190-194	192	2
185 " " 190	187	4	184.5-189.5	187	4	185-189	187	4
180 " " 185	182	5	179.5-184.5	182	5	180-184	182	5
175 " " 180	177	8	174.5-179.5	177	8	175-179	177	8
170 " " 175	172	10	169.5-174.5	172	10	170-174	172	10
165 " " 170	167	6	164.5-169.5	167	6	165-169	167	6
160 " " 165	162	4	159.5-164.5	162	4	160-164	162	4
155 " " 160	157	4	154.5-159.5	157	4	155-159	157	4
150 " " 155	152	2	149.5-154.5	152	2	150-154	152	2
145 " " 150	147	3	144.5-149.5	147	3	145-149	147	3
140 " " 145	142	1	139.5-144.5	142	1	140-144	142	1
$N = 50$			$N = 50$			$N = 50$		

For the rapid tabulation of scores within their proper intervals, method (C), in spite of its lack of precision, is usually to be preferred either to (B) or (A). In (A) it is fairly easy, even when one is on guard, to let a score of 160, say, slip into the interval "155 up to 160," owing simply to the presence of 160 at the upper limit of the step. Method (B) is clumsy and time-consuming because of the need for writing .5 at the beginning and end of every step. Method (C), while best for tabulation, offers the difficulty that in calculating measures of central tend-

ency and variability one must constantly remember that the *expressed* step limits are not the *actual* step limits, that interval "140-144" begins at 139.5 (not 140) and ends at 144.5 (not 144). If this is clearly understood, however, method (C) is as accurate as either (B) or (A). It will be generally used throughout this chapter.

The scores grouped within any given interval in a frequency distribution are assumed to be spread evenly over the entire interval. This assumption is made whether the step length is three, five, or ten units. If, however, we wish to represent *all* of the scores within a given interval by some single value, the midpoint of the step-interval is taken to be the logical choice. For example, in the interval 175-179 [Table 2, method (C)] all of the eight scores upon this step are represented by the single value 177, the midpoint of the interval.* Why 177 is the midpoint of this step-interval is shown graphically below:



A simple rule for finding the midpoint of a step-interval is

$$\text{Midpoint} = \text{lower limit of step} + \frac{(\text{upper limit} - \text{lower limit})}{2}.$$

In our illustration, $174.5 + \frac{(179.5 - 174.5)}{2} = 177$. Since the

step length is five units, it follows that the midpoint must be 2.5 units from the *lower limit* of the step, i.e., $174.5 + 2.5$; or 2.5 units from the *upper limit* of the step, i.e., $179.5 - 2.5$.

It is often a question whether the midpoint is, in fact, a fair representative of *all* of the scores upon a given step-interval. Referring to Table 1, we find that of the 10 scores on step "170 up to 175" (midpoint 172), three (170, 171, 171) are below the midpoint; three (172, 172, 172) are on the midpoint; and four (173, 173, 173, 174) are above the midpoint. Of the five scores

* The same value (namely, 177) is, of course, the midpoint of the interval when methods (A) and (B) are used.

upon step "180 up to 185," three (180, 181, 181) are below the midpoint (182); and two (183, 184) are above. The single score of 197 upon step "195 up to 200" falls exactly on the midpoint. In these examples the midpoint represents quite adequately the scores within the given step-intervals; but it must be admitted that the balancing of scores above and below the midpoint is not always so nearly equal as it is here. When the data are scanty, or when the distribution is badly skewed (p. 115), there may be many more scores on one side of a midpoint than on the other. When this happens, the midpoint does not fairly represent *all* of the scores within the given interval.

The assumption that the midpoint is the most representative score within an interval holds best when the number of scores in the distribution is large, and when the step-intervals are not very broad. Even when neither of these conditions fully obtains, however, the midpoint assumption is not greatly in error and is the best that we can make. In the long run, about as many scores will fall above as below the various midpoint values; and lack of balance in one interval will usually be offset by the opposite condition in another.

Measures of central tendency (p. 17) and of variability (p. 33) calculated from data grouped into step-intervals of five units, say, will usually vary slightly from the same measures calculated from the same data when ungrouped, or when grouped into step-intervals of, say, three or ten units. These variations arise from (1) differences in the size of the groups in which the data are classified, and (2) the fact that each score within a step-interval is assigned the value of the middle of the step instead of its actual value. Corrections are sometimes applied to the measures of variability to correct the *grouping error* thus introduced (p. 43). But usually the error which results from grouping is so slight that it may be neglected in ordinary statistical work.

III. STANDARDS OF ACCURACY IN COMPUTATION *

"How many places" to carry numerical results is a question which arises persistently in statistical computation. Sometimes a student, by discarding decimals, throws away legitimate data. More often, however, he tends to retain too many decimals, a practice which may give a false appearance of great precision not always justified by the original material.

In this section we shall give some of the generally accepted principles which apply to statistical calculation. Observance of these rules will lead to greater uniformity and reliability in calculation. They should be followed carefully in solving the problems given in this book.

1. Rounded Numbers

In calculation, numbers are usually "rounded" off to the standard of accuracy demanded by the problem. If we round off 8.6354 to two decimals it becomes 8.64; to one decimal, 8.6; to the nearest integer, 9. Measures of central tendency and variability, coefficients of correlation, and other measures, are rarely reported to more than two decimal places. Thus a mean of 52.6872 is usually reported as 52.69; a standard deviation of 12.3841 as 12.38; and a coefficient of correlation of .6278 as .63, etc. It is very doubtful whether much of the work in mental measurement warrants accuracy beyond the second decimal.

2. Significant Figures

The measurement 64.3 inches is assumed to be correct to the nearest tenth of an inch, its true value lying somewhere between 64.25 and 64.35 inches. Two places to the left of the decimal point, and one to the right are fixed, and hence 64.3 is said to contain *three* significant figures. The numbers 643 and .643 also contain three significant figures each.

In the number .003046 there are *four* significant figures, 3, 0, 4, and 6, the first two zeros serving merely to locate the deci-

* This section should be reviewed frequently, and referred to in solving the problems given in succeeding chapters

mal point. When used to locate a decimal point only, a zero is not considered to be a significant figure; .004, for example, has only *one* significant figure, the two zeros simply fixing the position of 4, the significant digit. The following illustrations should make clear the matter of significant figures:

- 136 has *three* significant figures.
- 136000 has *three* significant figures also. The true value of this number lies between 136,500 and 135,500. Only the first three digits are definitely fixed, the zeros serving simply to locate the decimal point or fix the size of the number.
- 1360. has *four* significant figures; the decimal indicates that the zero in the fourth place is known — and hence significant.
- .136 has *three* significant figures.
- .1360 has *four* significant figures; the zero fixes the fourth place.
- .00136 has *three* significant figures; the first two zeros merely locate the decimal point
- 2.00136 has *six* significant figures; the integer, 2, makes the two zeros to the right of the decimal point significant.

3. Exact and Approximate Numbers

It is necessary in calculation to make a distinction between *exact* and *approximate* numbers. An exact number is one which is found by counting: 10 children, 150 test scores, 20 desks are examples. Approximate numbers result from the measurement of variable quantities. Test scores, for example, are approximate numbers since they represent intervals on some scale. Thus a score of 61 means 60.5–61.5 (p. 2) and a measured height of 47.5 inches means 47.45–47.55 inches. Calculations with exact numbers may, in general, be carried to as many decimals as we please since we may assume as many significant figures as we wish. For example, 110 test scores, which means that exactly 110 subjects were tested, could be written $N = 110.000 \dots$ to n significant figures. Calculations based upon approximate numbers, however, depend upon, and are limited by, the number of significant figures in the numbers which enter into the calculations. This will be clearer in the following “rules.”

4. Rules for Computation**(1) Accuracy of a Product**

(a) The number of significant figures in the product of two or more approximate numbers will equal the number of significant figures in that one of the numbers which is the least accurate, i.e., contains the smallest number of significant figures. To illustrate:

$125.5 \times 7.0 = 880$ not 878.5, because 7.0, the less accurate of the two numbers, contains only two significant figures. The number 125.5 contains four significant figures.
 $125.5 \times 7.000 = 878.5$. Both numbers now contain four significant figures; hence their product also contains four significant figures.

(b) When multiplying an exact number by an approximate number, the number of significant figures in the product is determined by the number of significant figures in the approximate number. To illustrate:

If each of 12 children (12 is an exact number) has an M.A. of 8 years (8 is an approximate number) the product 12×8 must be written either as 90 or 100, since the approximate number has only *one* significant digit. If, however, each M.A. of 8 years may be written as 8.0, the product 12×8.0 can be written as 96, since 8.0 contains *two* significant digits.

(2) Accuracy of a Quotient

(a) When dividing one approximate number by another approximate number, the number of significant figures in the quotient will equal the number of significant figures in that one of the two numbers (dividend or divisor) which is the less accurate, i.e., has the smaller number of significant digits. Illustrations:

$\frac{9.27}{41}$ should be written .23, not .22609, since 41 (the less accurate number) contains only two significant figures.

$\frac{16}{4724}$ should be written .0034, not .0033869, since 16 (the less accurate number) has two significant figures.

(b) In dividing an approximate number by an exact number, the number of significant figures in the quotient will equal the number of significant figures in the approximate number. Illustrations:

$\frac{9.27}{41}$ should be written .226, since 9.27, the approximate number, has three significant figures. The number 41 is an exact number.

$\frac{8541}{50}$ should be written 170.8, not 170.82, since 8541, the approximate number, contains only four significant figures.

(c) In dealing with exact numbers, quotients may be written to as many decimals as one wishes.

(3) Accuracy of a Root or Power

(a) The square root of an approximate number can contain no more significant figures than there are in the number itself. The number of significant figures retained in a square root is usually less (often one-half) than the number of significant figures in the number. For example, $\sqrt{159.5600}$ is usually written 12.63, and not 12.63176, although the original number (159.5600) contains seven significant figures.

(b) The square, or higher power, of an approximate number contains as many significant figures as there are in the original number (and no more). For example, $(.034)^2 = .0012$ (two significant figures) and not .001156 (four significant figures).

(c) Roots and powers of exact numbers may be taken to as many decimal places as one wishes.

(4) Accuracy of a Sum or Difference

The number of decimal places to be retained in a sum or difference should be no greater than the number of decimals in the least accurate of the numbers added or subtracted. Illustrations:

$362.2 + 18.225 + 5.3062 = 385.7$ not 385.7312, since the least accurate number (362.2) contains only four significant figures.

$362.2 - 18.245 = 344.0$, not 343.955, since the less accurate number (362.2) contains only four significant figures.

PROBLEMS

1. Indicate which of the following variables fall into continuous and which into discrete series: (a) time; (b) salaries in a large business firm; (c) sizes of elementary school classes; (d) age; (e) census data; (f) distance travelled by car; (g) football scores; (h) weight; (i) numbers of pages in 100 books; (j) mental ages.
2. Write the upper and lower limits of the following scores in accordance with the two definitions of a score in continuous series, given on pages 2 and 3:

62	175	1
8	312	87

3. Suppose that sets of scores have the ranges given below. Indicate how large an interval, and how many intervals, you would use in drawing up a frequency distribution of each set.

Range	Size of Interval	Number of Intervals
16 to 87		
0 to 46		
110 to 212		
63 to 151		
4 to 12		

4. In each of the following cases write the lower and upper limits of the given step-intervals (following the first definition of a score, given on page 2) and the midpoint of each interval.

45-47	162.5-167.5	63-67
1-4	80 up to 90	16-17

- (a) Tabulate the following 25 scores into two frequency distributions, using (1) an interval of 3, and (2) an interval of 5 units. Let the first interval begin with 60.

72,	75	77	67	72
81	78	65	86*	73
67	82	76	76	70
83	71	63	72	72
* 61	67	84	69	64

- (b) The following 100 scores were made on the Thorndike Intelligence Examination for High School Graduates by applicants

for admission to Columbia College.* Tabulate these scores into three frequency distributions, using step-intervals of 3, 5, and 10 units.

63	80	75	90	81	83
78	81	83	83	89	98
46	90	103	81	71	93
82	78	86	85	73	83
74	86	84	72	63	76
103	78	85	81	105	94
78	101	76	96	74	75
86	65	80	81	98	56
103	90	92	85	78	73
87	75	102	58	78	95
73	73	73	96	83	110
95	90	87	86	96	98
82	86	70	70	95	71
89	86	85	72	94	92
73	84	79	74	88	72
92	86	93	84	50	
85	76	82	99	91	

3. (a) Round off the following numbers to two decimals:

3.5872	74.168
46.9223	25.393

- (b) How many significant figures in each of the following:

.00046	91.00	1.03
46 02	18.365	15.0048

- (c) Write the answers to the following:

$127.4 \times .0036 =$	(both numbers approximate)
$200.0 \div 5.63 =$	" " "
$62 \times .053 =$	(first number exact, second approximate)
$364.2 + 61.596 =$	(both numbers approximate)
$364.2 - 61.596 =$	" " "
$\sqrt{4786} =$	
$(18.6)^2 =$	

* Sommerville, R. C., *Physical, Motor, and Sensory Traits*, Archives of Psychology, 1924, 75, pp. 101-103.

ANSWERS

2. 61.5-62.5 and 62-63; 174.5-175.5 and 175-176;
 7.5- 8.5 and 8-9; 311.5-312.5 and 312-313;

.5- 1.5 and 1-2
 86.5-87.5 and 87-88

3.	Size of Interval	No of Intervals
	5	15
	3 or 4 or 5	16 or 12 or 10
	10	11
	5 or 10	18 or 9
	1	<u>9</u>
		Midpoint
4.	44.5- 47.5	46.0
	.5- 4.5	2.5
	162.5-167.5	165.0
	79.5- 89.5	84.5
	62.5- 67.5	65.0
	15.5- 17.5	16.5
6. (a)	3.59	74.17
	46.92	25.19
(b)	2 4	3
	4 5	6
(c)	.46	
	35.5	
	3.3	
	425.8	
	302.6	
	6.918	
	346	

CHAPTER II

MEASURES OF CENTRAL TENDENCY

WHEN scores or other measures have been tabulated into a frequency distribution, as shown in Chapter I, our next task, generally, is to calculate one or more measures of *central tendency*. The value of a measure of central tendency is twofold. First, it is a single measure which represents *all* of the scores made by the group, and as such gives a concise description of the performance of the group as a whole; and second, it enables us to compare two or more groups in terms of typical performance. There are three "averages" or measures of central tendency in common use, (1) the *arithmetic mean*, (2) the *median*, and (3) the *mode*. We shall consider the calculation of these three measures in order.

I. CALCULATION OF MEASURES OF CENTRAL TENDENCY

1. The Arithmetic Mean or "Average" (M)

The arithmetic mean or average* is the best known measure of central tendency. It may be defined simply as the sum of the separate scores or other measures in a series divided by their number. To illustrate: if a man earns \$3, \$4, \$3.50, \$5, and \$4.50 on five successive days his mean daily wage (\$4.00) is obtained by dividing the sum of his daily earnings by the number of days he has worked. The formula for the arithmetic mean (M) of a series of ungrouped measures is

$$M = \frac{\sum X}{N} \quad (1)$$

(*arithmetic mean calculated from ungrouped measures*)

in which N is the number of measures in the series, X stands

* Popularly, the average is the term used for the arithmetic mean. In statistical work, however, the term average is often used as a general expression to cover any measure of central tendency.

for scores or other measures, and the symbol Σ means "sum of."

When measures have been grouped into a frequency distribution, the arithmetic mean is calculated by a slightly different method from the one given above. The two illustrations given in Table 3 will make the differences clear. The first example shows the calculation of the mean of the 50 Army Alpha scores which were tabulated into a frequency distribution in Table 1. Note that we first calculate the fX column by multiplying the midpoint (X) of each step-interval by the number of scores (f) on it; and that the mean (170.80) is then simply the sum of the fX (8540) * divided by N (50). The use of the midpoint for all of the scores within an interval is made necessary by the fact that scores grouped into steps lose their identity and must thereafter be represented by the midpoint of the particular interval in which they fall. Hence, we multiply or "weight" the midpoint of each step by the frequency upon that step; add the fX and divide by N to obtain the mean. The formula may be written

$$M = \frac{\Sigma fX}{N} \quad (2)$$

(arithmetic mean calculated from measures grouped into a frequency distribution)

The second example in Table 3 is another illustration of the calculation of the mean from grouped data. This frequency distribution represents 200 scores made by a group of adults upon a Cancellation Test. Scores have been classified by method (B), page 6, into nine steps; and since the step-interval is four units, the midpoint of each step is found by adding $\frac{1}{2}$ of 4 to the beginning of each step. For example, $103.5 + 2.0 = 105.5$. The fX column (found as shown above) totals 23,888.0; and N equals 200. Hence, applying formula (2), the arithmetic mean is found to be 119.44 (to two decimals).

* The sum 8540 may be written 8540.000 . . . (i.e., to any number of significant figures) since each midpoint value (X) is an exact point on the scale, and the f 's are also exact numbers. The mean (170.80), however, has been carried only to two decimals — the usual standard of accuracy for measures of central tendency.

TABLE 3

THE CALCULATION OF THE MEAN, MEDIAN, AND MODE
FROM DATA GROUPED INTO A FREQUENCY
DISTRIBUTION

1. Data from Table 1, 50 Army Alpha Scores
Step-interval = 5

Step-Intervals Scores	Midpoint \bar{X}	f	fX
195-199	197	1	197
190-194	192	2	384
185-189	187	4	748
180-184	182	5	910
175-179	177	8	1416
170-174	172	10	1720
165-169	167	6	1002
160-164	162	4	648
155-159	157	4	628
150-154	152	2	304
145-149	147	3	441
140-144	142	1	142
		$N = 50$	8540
		$N/2 = 25$	

$$(1) \text{ Mean} = \frac{\Sigma fX}{N} = \frac{8540}{50} = 170.80$$

$$(2) \text{ Median} = 169.5 + \frac{5}{10} \times 5 = 172.00$$

(3) Mode (crude) falls on step-interval 170-174 or at 172

2. Scores Made by 200 Adults upon a Cancellation Test
Step-interval = 4

Step-Intervals Scores	Midpoint \bar{X}	f	fX
135.5-139.5	137.5	3	412.5
131.5-135.5	133.5	5	667.5
127.5-131.5	129.5	16	2072.0
123.5-127.5	125.5	23	2886.5
119.5-123.5	121.5	52	6318.0
115.5-119.5	117.5	49	5757.5
111.5-115.5	113.5	27	3064.5
107.5-111.5	109.5	18	1971.0
103.5-107.5	105.5	7	738.5
		$N = 200$	23888.0
		$N/2 = 100$	

$$(1) \text{ Mean} = \frac{\Sigma fX}{N} = \frac{23,888.0}{200} = 119.44$$

$$(2) \text{ Median} = 115.5 + \frac{4}{8} \times 4 = 119.42$$

(3) Mode (crude) falls on step-interval 119.5-123.5 or at 121.5

In both of the illustrations in Table 3, the M of the scores made by the members of a *group* was found. However, we may use either formula (1) or formula (2) to calculate the M of a number of measurements made upon the same individual. If an individual's reaction time to light is measured 100 times, and the measures tabulated into a frequency distribution, the M may be found in exactly the same way in which we find the "average" reaction time to light of 100 different observers.

2. The Median (Mdn)

(1) The Calculation of the Median when Scores are Ungrouped

When ungrouped scores or other measures are arranged in order of size, the median is the *midscore* or midpoint in the series. Two situations arise in the computation of the median from ungrouped data: (1) when N is odd, and (2) when N is even. To consider, first, the case where N is odd, suppose we have the following "mental ages" — 7, 10, 8, 12, 9, 11, 7, calculated from seven performance tests. If we arrange these seven scores in order of size

7 7 8 (9) 10 11 12

the median score (or midscore) is 9, since 9 is the middle number in the series. There are three scores above 9 and three below.

Now suppose we drop off the first score of 7 so that our series becomes

7 8 9 \uparrow 10 11 12

Since $N(6)$ is now even, there is clearly no median *score*. We may, however, take the median arbitrarily as the point midway between the two middlemost scores. The two middle scores are 9 and 10; and the Mdn lies at 9.5 or halfway between these two.

The formula for finding the median of a series of ungrouped scores is

$$\text{Median} = \text{the } \frac{(N + 1)}{2}\text{-th score in order of size} \quad (3)$$

(*median score from ungrouped data*)

In our first illustration above, the median is the $\frac{(7 + 1)}{2}$ or fourth score counting in from either end of the series, that is, 9. In our second illustration, the median is the $\frac{(6 + 1)}{2}$ or 3.5th score in order of size, that is, 9.5.

(2) The Calculation of the Median when Scores are Grouped into a Frequency Distribution

The method of computing the median when data are grouped into a frequency distribution differs in several respects from the method described above for ungrouped scores. When scores in a continuous series are grouped into a frequency distribution, the median is defined as the 50% point in the distribution. To locate the median, therefore, we take 50% (i.e., $N/2$) of our scores, and count halfway into the distribution. The method is illustrated in the two examples in Table 3. Since there are 50 scores in the first distribution, $N/2 = 25$, and the median is that point in our distribution of Army Alpha scores which has 25 scores on each side of it. Beginning at the small-score end of the distribution, and adding up the scores in order, we find that steps 140-144 to 165-169, inclusive, contain just 20 f 's — five scores short of the 25 necessary to locate the median. The next step, 170-174, contains ten scores assumed to be spread evenly over the step-interval (p. 8). In order to get the five extra scores needed to make exactly 25, therefore, we take $5/10 \times 5$ (the step length) and add this amount (2.5) to 169.5, the beginning of the step-interval 170-174. This puts the Mdn at $169.5 + 2.5$ or at 172.0. Note carefully that in a frequency distribution the median like the mean is a *point* and not a *score*.

A second illustration of the calculation of the median from data grouped into a frequency distribution is given in Table 3 (2). There are 200 scores in this distribution; hence, $N/2 = 100$, and the median must lie at a point 100 scores-distant from either end of the distribution. If we begin at the small-

score end of the distribution (103.5–107.5) and add the scores in order, 52 scores take us *through* step 111.5–115.5. The 49 scores on the next step-interval (115.5–119.5) plus the 52 already counted off total 101 — *one* score too many to give us 100, the point at which the median falls. To get the 48 scores needed to make *exactly* 100 we must take $48/49 \times 4$ (the step length) and add this amount (3.92) to 115.5, the beginning of the step-interval 115.5–119.5. This procedure takes us exactly 100 scores into the distribution, and locates the median at 119.42.

A formula for calculating the *Mdn* when the data have been classified into a frequency distribution is

$$Mdn = l + \left(\frac{\frac{N}{2} - F}{f_m} \right) i \quad \checkmark \quad (4)$$

(median computed from data grouped into a frequency distribution)

where

l = lower limit of the step-interval upon which the median lies

$\frac{N}{2}$ = one-half the total number of scores

F = sum of the scores on all steps *below* l

f_m = frequency (number of scores) *within* the step upon which the median falls

i = length of the step-interval.

To illustrate the use of formula (4), consider the first example in Table 3. Here $l = 169.5$, $N/2 = 25$, $F = 20$, $f_m = 10$, and $i = 5$. Hence, the median falls at $169.5 + \frac{(25 - 20)}{10} \times 5$ or at 172.0. In the second example, $l = 115.5$, $N/2 = 100$, $F = 52$, $f_m = 49$, and $i = 4$. The median, therefore, is $115.5 + \frac{(100 - 52)}{49} \times 4$ or 119.42.

We may summarize the steps involved in computing the *Mdn* from data tabulated into a frequency distribution as follows:

- (1) Find $N/2$
- (2) Begin at the small-score end of the distribution and count off the scores in order (i.e., the F) up to the lower limit (l) of the step-interval which contains the median.
- (3) Compute the number of scores necessary to fill out $N/2$, i.e., compute $N/2 - F$. Divide this quantity by the number of scores on the step-interval containing the median (f_m); and multiply the result by the step length (z).
- (4) Add the amount obtained by the calculations in (3) to the lower limit (l) of the step which contains the Mdn . This will give the median of the distribution.

- (3) Calculation of the Mdn when (a) the Frequency Distribution contains Gaps; and when (b) the first or last Step-interval has Indeterminate Limits

(a) Difficulty often arises when it becomes necessary to calculate a median from a distribution of scores in which there are gaps, or zero frequency upon some steps. A method of computing the median in such situations is shown in Table 4. Since $N = 10$, and $N/2 = 5$, we count *up* the frequency column five scores through step 6-7. Ordinarily, this would put the median at 7.5, the beginning of step 8-9. If we check this median value, however, by counting *down* the frequency column five scores, the median falls at 11.5, the beginning of step 12-13. Obviously, the discrepancy between these two values of the median is due to the two step-intervals 8-9 and 10-11 (each of which has zero frequency) which lie between step 6-7 and step 12-13. In order to have the median come out at the same point, whether computed from the top or the bottom of the frequency distribution, the procedure usually followed in cases like this is to have step 6-7 *include* 8-9, thus becoming 6-9; and to have step 12-13 *include* 10-11, becoming 10-13. Lengthening the step-interval for these two steps from two to four units eliminates the zero frequency on the adjacent steps by spreading the frequency on these steps over it. If now we count off five scores, going *up* the frequency column through step 6-9,

TABLE 4
COMPUTATION OF THE MEDIAN WHEN THERE ARE GAPS
IN THE DISTRIBUTION

Step-Intervals Scores	<i>f</i>	
20-21	2	
18-19	1	
16-17	0	
14-15	0	
✓ 12-13	2	} 10-13
10-11	0	
8-9	0	} 6-9
- 6-7	2	
- 4-5	1	
2-3	1	
0-1	1	
	$N = 10$	
	$N/2 = 5$	

$$Mdn = 9.5 + \frac{1}{2} \times 2 = 9.5$$

the median falls at 9.5, the upper limit of this step, or the lower limit of the next step, 10-13. Also, counting *down* the frequency column five scores, we arrive at a median value of 9.5, the beginning, or lower limit, of step 10-13. Computation from the two ends of the series now gives consistent results — the median is 9.5 in both instances.

(b) When scores scatter widely, the last step-interval in a frequency distribution may be designated as "80 and above" or simply as 80+. This description indicates that *all* scores ~~above~~ 80 are simply thrown into this step-interval, the upper limit of which is thus indeterminate. The same lumping together of scores may also occur at the beginning of the distribution, as, for example, when the first interval is designated "20 and below" or 20-. The lower limit of the beginning step-interval is now indeterminate. In irregular distributions like these, the median is readily computed since each score is simply counted as one frequency whether accurately classified or not. But it is difficult to calculate a mean value which is a representative measure of central tendency when the value of the midpoint of one or more intervals is uncertain. Since the

mean depends upon the absolute size of the scores (or their midpoints) it is directly affected by indeterminate step limits.

Another case may be mentioned in the present connection in which the median is a better measure of central tendency than the mean. This is when there are many very high scores or many zero scores. Such scores count simply as *single* frequencies in determining the median; but they depress and distort the mean when their absolute values are used in its calculation.

3. The Mode (Range like Score)

In a simple ungrouped series of scores the "crude" or "empirical" mode is that single measure or score which occurs most frequently. For example, in the series 10, 11, 11, 12, 12, 13, 13, 13, 14, 14, the most often recurring measure, namely 13, is the crude or empirical mode. When data are grouped into a frequency distribution, the crude mode is usually taken to be the midpoint of that step-interval which contains the greatest frequency. In example 1, Table 3, the interval 170-174 contains the greatest frequency and hence 172, its midpoint, is the crude mode. In example 2, Table 3, the greatest frequency falls on step 119.5-123.5 and the crude mode is at 121.5, the midpoint of the step.

When calculating the mode from a frequency distribution, we distinguish between the "true" mode and the crude mode. The true mode is the point (or "peak") of greatest concentration in the distribution; that is, the point at which more measures fall than at any other point. When the scale is divided into finely graduated units, when scores are recorded exactly, and when N is large, the crude mode closely approaches the true mode. Ordinarily, however, the crude mode is only approximately equal to the true mode. A formula for calculating the true mode, when the frequency distribution is symmetrical, or at least not badly skewed (p. 115), is

$$\text{Mode} = 3 \text{ Mdn} - 2 \text{ Mean} \quad (5)$$

(approximation to the true mode calculated from a frequency distribution)

If we apply this formula to the data in Table 3, the mode is 174.40 for the first distribution, and 119.38 for the second. The first mode is somewhat larger and the second slightly smaller than the crude modes obtained from the same distributions.

The crude mode is often an unstable measure of central tendency. This instability, however, is not so serious a drawback to the usefulness of the crude mode as might seem at first glance. When the mode is employed as a simple, inspectional "average," that is, to indicate in a rough way the center of concentration in the distribution, it need not be calculated as exactly as the median or mean.

II. CALCULATION OF THE MEAN BY THE "ASSUMED MEAN" OR SHORT METHOD

In Table 3 the mean was calculated by multiplying the midpoint (X) of each step-interval by the frequency (number of scores) on the step, summing up these values (the fX column) and dividing by N , the number of scores. This straightforward or Long Method gives accurate results but it often requires the handling of large numbers and entails tedious calculation. Because of this fact, the "Assumed Mean" method, or simply the Short Method, has been devised for computing the mean. The Short Method does not apply to the calculation of the median or the mode. These measures are always found by methods with which we are now familiar.

The most important fact to remember in calculating the mean by the Short Method is that we "guess" or "assume" a mean at the outset, and later apply a correction to this assumed mean (AM) in order to obtain the actual mean (M) [see Table 5]. There is no set rule for guessing a mean.* The best plan is to take the midpoint of a step somewhere near the center of the distribution; and if possible the midpoint of that step-interval which contains the greatest frequency. In Table 5, the

* The method outlined here gives accurate results no matter where the mean is assumed.

greatest f is on step 170-174, which also happens to be near the center of the distribution. Hence the AM is taken at 172, the middle of this step. The question of the AM settled, we next determine the correction which must be applied to the AM in order to get M in the following way:

- (1) First we fill in the x' column,* column 4. Here are entered the deviations of the midpoints of the different steps measured from the AM in units of step-interval. Thus 177, the midpoint of step 175-179, deviates from 172, the AM , by one step-interval; and a "1" is placed in the x' column opposite 177. In like manner, 182 deviates two steps from 172; and a "2" goes in the x' column opposite 182. Reading on up the x' column from 172, we find the succeeding entries in the same way to be 3, 4, and 5. The last entry, 5, is the step-deviation of 197 from 172; the actual score-deviation, of course, is 25.

Returning to 172, we find that the x' of this score measured from the AM (from itself) is zero; hence a zero is placed in the x' column opposite step 170-174. Below 172, all of the x' entries are negative, since all of the midpoints are less than 172, the AM . So the x' of 167 from 172 is - 1 step-interval; and the x' of 162 from 172 is - 2 step-intervals. The other x' 's are - 3, - 4, - 5, and - 6.

- (2) The x' column completed, we next compute the fx' column, column 5. The fx' entries are found in exactly the same way as in the Long Method (p. 18). Each x' in column 4 is multiplied or "weighted" by the appropriate f in column 3. Note again that in the Short Method we multiply each x' by its deviation from the AM in units of step-interval, instead of by its actual deviation from the mean of the distribution. For this reason, the computation of the fx' column is much more simple than is the calculation of the fx column in the Long Method. All of the fx' above (greater than) the

* x' is regularly used to denote the deviation of a score X from the assumed mean (AM); x is the deviation of a score X from the actual mean (M) of the distribution.

TABLE 5

THE CALCULATION OF THE MEAN BY THE SHORT METHOD
(Data from Table 1, 50 Army Alpha Scores)

(1) Step-Intervals Scores	(2) Midpoint \bar{X}	(3) f	(4) x'	(5) fx'
195-199	197	1	5	5
190-194	192	2	4	8
185-189	187	4	3	12
180-184	182	5	2	10
175-179	177	3	1	3
170-174	172	10	0	0
165-169	167	6	-1	-6
160-164	162	4	-2	-8
155-159	157	4	-3	-12
150-154	152	2	-4	-8
145-149	147	3	-5	-15
140-144	142	1	-6	-6
		$N = 50$		$- 55$
$AM = 172.00$		$c = -\frac{12}{50} = -.240$		
$ci = -1.20$		$i = 5$		
$M = 170.80$		$ci = -.240 \times 5 = -1.20$		

AM are positive; and all *below* (smaller than) the AM are negative, since the signs of the fx' depend upon the signs of the x' .

- (3) From the fx' column the correction is obtained as follows. The sum of the positive values in the fx' column is 43; and the sum of the negative values in the fx' column is 55. There are, therefore, 12 more *minus* fx' values than *plus* (the algebraic sum is -12); and -12 divided by 50 (N) gives $-.240$ which is the correction (c) in *units of step-interval*. If we multiply c ($-.240$) by i (5), the length of the step, the result is ci (-1.20) the score correction, or the correction in *score units*. When -1.20 is added to 172.00, the AM , the result is the actual mean, 170.80.

We may summarize the process of calculating the mean by the Short Method as follows:

- (1) Organize the scores or measures into a frequency distribution.
- (2) "Guess" or "assume" a mean as near the center of the

distribution as possible, and preferably on the step containing the greatest frequency.

- (3) Find the deviation of the midpoint of each step-interval from the AM in units of step-interval.
- (4) Multiply or weight each step-deviation (x') by its appropriate f — the f opposite it.
- (5) Find the algebraic sum of the plus and minus fx' and divide this sum by N , the number of cases. This gives c , the correction in units of step-interval.
- (6) Multiply c by the step length (i) to get ci , the score correction.
- (7) Add ci algebraically to the AM to get the actual mean. Sometimes ci will be positive and sometimes negative, depending upon where the mean has been assumed. The method works equally well in either case.

III. WHEN TO USE THE VARIOUS MEASURES OF CENTRAL TENDENCY

The beginning student of statistics is often puzzled to know which measure of central tendency to use in a given problem. The following summary will serve as a convenient guide for most statistical work.

1. *Use the mean*

- (1) When each score or measure should have equal weight in determining the central tendency.
- (2) When the measure of central tendency having the highest reliability is desired.
- (3) When product-moment coefficients of correlation are to be subsequently computed, or the differences between the mean abilities of groups compared.

2. *Use the median*

- (1) When a quick and easily computed measure of central tendency is wanted.
- (2) When there are extreme measures which would affect the mean disproportionately (p. 25).

- (3) When it is desired that certain scores should influence the central tendency but all that is known about them is that they are above or below the median (p. 24).

3. *Use the mode*

- (1) When the most often recurring score is sought.
 (2) When a quick approximate measure of concentration is all that is wanted.

PROBLEMS

1. Calculate the mean, median, and mode for the following frequency distributions. Use the Short Method in computing the mean.

1) Scores	f	x'	$\sum fx'$	(2) Scores	f	d	fd
70-71	2	+5	+10	90-94	2	5	
68-69	2	+4	+8	85-89	2	4	
66-67	3	+3	+9	80-84	4	3	
64-65	4	+2	+8	75-79	8	2	
62-63	6	+1	+6	70-74	6	1	
60-61	7	0	0	65-69	11	0	
58-59	5	-1	-5	60-64	9	-1	
56-57	4	-2	-8	55-59	7	-2	
54-55	2	-3	-6	50-54	5	-3	
52-53	3	-4	-12	45-49	0	-4	
50-51	1	-5	-5	40-44	2	-5	
	$N = 39$		$\sum fx' = 36$		$N = 56$		

1) Scores	f	x'	$\sum fx'$	(4) Scores	f
120-122	2	+		100-109	5
117-119	2	+		90-99	9
114-116	2	+		80-89	14
111-113	4	+		70-79	19
108-110	5	+		60-69	21
105-107	9	+		50-59	30
102-104	6	+		40-49	25
99-101	3	-		30-39	15
96-98	4	-		20-29	10
93-95	2	-		10-19	8
90-92	1	-		0-9	6
	$N = 40$				$N = 162$

2. Compute the mean and the median for each of the two distributions in problem 5(a), page 14, tabulated in 3- and 5-unit step-intervals. Compare the two means and the two medians, and explain any discrepancy found. (Let the first step in the first distribution be 61-63; the first step in the second distribution, 60-64.)
3. (a) Compute the median of the following 16 scores

Scores	<i>f</i>
20 to 22	2
18 " 20	2
16 " 18	0
14 " 16	4
12 " 14	0
<u>10 " 12</u>	0
8 " 10	4
6 " 8	0
4 " 6	0
2 " 4	0
0 " 2	4
$N = 16$	

- (b) In a group of 50 children, the eight children who took longer than five minutes to complete a performance test were marked D.N.C. (did not complete). In computing a measure of central tendency for this distribution of scores, what measure would you use, and why?
- (c) Find the medians (i.e., midscores) of the following arrays of ungrouped scores:
- (1) 21, 24, 27, 29, 29, 30, 32, 33, 35, 38, 42, 45.
- (2) 54, 59, 64, 67, 70, 72, 73, 75, 78, 83, 90
4. The time by your watch is 10:31 o'clock. In checking this reading with two friends, you find that their watches give the time as 10:25 and 10:34. Assuming that the three watches are equally good timepieces, what do you think is the most probable "correct time"?
5. What is meant popularly by the term "law of averages"?
6. (a) When one uses the term "in the mode" does he have reference to the mode of a distribution?
- (b) What is approximately the modal time for each of the following meals: breakfast, lunch, dinner. Explain your answers.

ANSWERS

1. (1) Mean = 60.76
Median = 60.76
Mode = 60.86
- (2) Mean = 67.36
Median = 66.77
Mode = 65.59
- (3) Mean = 106.00
Median = 105.83
Mode = 105.49
- (4) Mean = 55.43
Median = 55.17
Mode = 54.65
2. Step-interval = 3
Mean = 72.92
Median = 71.75
- Step-interval = 5
Mean = 73.00
Median = 72.71
3. (a) Median = 11.5
(c) (1) Median = 31
(2) Median = 72
4. Mean is 10.30.

CHAPTER III

MEASURES OF VARIABILITY

In Chapter II we discussed the calculation of measures of central tendency — measures typical or representative of our set of scores as a whole. Ordinarily, the next step is to find some measure of the *variability* of our scores, i.e., of the “scatter” or “spread” of the separate scores or measures around their central tendency. It will be the task of this chapter to show how the measures of variability are calculated.

The usefulness of a measure of variability can be shown by a simple example. Suppose we have given a test of controlled association to a group of 50 boys and to a group of 50 girls. The mean scores are, boys, 34.6 secs., and girls, 34.5 secs. So far as the means go there is no difference in the performance of the two groups. But suppose, upon examining the original data, we find that the boys' scores range from 15 to 51 secs. and the girls' scores from 19 to 45 secs. This fact would make it evident at once that in a general way the boys “cover more territory,” are more *variable*, than the girls; and this greater variability may be of more interest than the lack of difference between the mean scores. If a group is *homogeneous*, that is, made up of individuals of nearly the same ability, most of the scores will fall near the same point on the scale, the range will be relatively short, and the variability will be small. But if the group contains individuals of widely differing capacities, scores will be strung out from high to low, the range will be relatively wide, and the variability large. Four measures have been devised to take account of the variability within a set of measures. These are (1) the *range*, (2) the *quartile deviation* or *Q*, (3) the *average deviation* or *AD*, and (4) the *standard deviation* or *SD*. In addition to these four measures of absolute variability, a

measure of relative variability is often employed in comparing different distributions. This *coefficient of variation*, V , will be discussed in a later section (p. 51).

I. CALCULATION OF MEASURES OF VARIABILITY

1. The Range

In grouping the scores in Table 1 into a frequency distribution (p. 5) we have already had occasion to use *the range*. It may be redefined simply as the interval between the largest and the smallest scores. In the illustration above, the range of the boys' scores was 51-15 or 36 secs. and the range of the girls' scores 45-19 or 26 secs. The range is the most general measure of spread or scatter. It includes the scale distance covered and is employed when we wish to make a rough comparison of two or more groups for variability; or when the number of scores is too small to justify the calculation of a more precise measure. The range takes account of the extremes of the series only; and hence is unreliable when many or large gaps occur in the frequency distribution.

2. The Quartile Deviation or Q

The quartile deviation or Q is one-half of the distance between the 75th and 25th percentile points in a frequency distribution. The 25th percentile, called Q_1 , is the *first* quarter or *quartile* point on the scale, the point below which lie 25% of the measures. The 75th percentile, or Q_3 , is the *third* quarter or *quartile* point on the scale, the point below which lie 75% of the measures.*

In order to find Q , we must first calculate the 75th and 25th percentiles. These values are found by exactly the same method as that employed in calculating the median. To find Q_1 we count off 25% of the scores from the beginning of the distribution (low end); and to find Q_3 we count off 75% of the scores from the low end of the distribution, or 25% from the high end.†

Table 6 illustrates the calculation of Q for the distribution of

* It may be noted that the second quartile point, Q_2 , is the median.

† Q_1 and Q_3 may also be calculated from the formula given on page 76.

50 Alpha scores tabulated in Table 1. To find Q_1 , we count off $\frac{1}{4}$ of the total number of scores (12.5) from the *low* score end of the distribution. When the scores (f) are added in order, the first four step-intervals (the steps 140-144 to 155-159, inclusive) are found to contain 10 scores. The next step, 160-164, contains four scores. These scores are assumed to be spread evenly over the entire step (p. 8). We need only 2.5 additional scores to make up the necessary 12.5; hence we take $\frac{2.5}{4} \times 5$ (the step length) and add this amount, 3.13, to 159.50, the beginning of the step (p. 2). This calculation locates Q_1 at 162.63.

We find Q_3 in the same way by counting off $\frac{3}{4}$ of the scores (37.5) from the small score end of the distribution. The f 's on steps 140-144 to 170-174, inclusive, added in order, total 30. The next step, 175-179, contains eight scores. To round out the necessary 37.5, therefore, we take $\frac{7.5}{8} \times 5$ (step length) and add this amount (4.69) to 174.50, the beginning of the step. This puts Q_3 at 179.19.

When Q_1 and Q_3 are known, the quartile deviation is calculated from the formula

$$Q = \frac{Q_3 - Q_1}{2} \quad (6)$$

(quartile deviation calculated from grouped data)

In the present problem, $Q = \frac{179.19 - 162.63}{2}$ or 8.28

A second illustration of the calculation of Q from a frequency distribution is given in Table 6 (2). Since the N of this distribution is 200, $\frac{1}{4}$ of the measures equals 50. The steps 103.5-107.5 and 107.5-111.5 contain 25 scores; and the next step, 111.5-115.5, contains 27 scores, which make a total of two more than the 50 wanted. To find the point reached by just 50 scores, we must take $\frac{25}{27} \times 4$ (the step length) and add this amount (3.70) to 111.50, the lower limit of step 111.5-115.5. This locates Q_1 at 115.20.

TABLE 6
THE CALCULATION OF THE Q , AD , AND SD FROM DATA GROUPED
INTO A FREQUENCY DISTRIBUTION

1 Data from Table 1, 50 Army Alpha Scores

(1) Step-Intervals Scores	(2) Midpoint X	(3) f	(4) x	(5) fx	(6) fx^2
195-199	197	1	26.20	26.20	686.44
190-194	192	2	21.20	42.40	898.88
185-189	187	4	16.20	64.80	1049.76
180-184	182	5	11.20	56.00	627.20
175-179	177	8	6.20	49.60	307.52
170-174	172	10	1.20	12.00	14.40
165-169	167	6	-3.80	-22.80	86.64
160-164	162	4	-8.80	-35.20	309.76
155-159	157	4	-13.80	-55.20	761.76
150-154	152	2	-18.80	-37.60	706.88
145-149	147	3	-23.80	-71.40	1699.32
140-144	142	1	-28.80	-28.80	829.44
		$N = 50$		502.00	7978.00

Mean = 170.80 (Table 3)

$$\frac{N}{4} = 12.5, \text{ therefore,}$$

$$\frac{3N}{4} = 37.5, \text{ therefore,}$$

$$Q_1 = 159.5 + \frac{2.5}{4} \times 5 = 162.63$$

$$Q_3 = 174.5 + \frac{7.5}{8} \times 5 = 179.19$$

$$Q = \frac{Q_3 - Q_1}{2} = \frac{179.19 - 162.63}{2} = 8.28$$

$$AD = \frac{\sum fx}{N} = \frac{502.00}{50} = 10.04$$

$$SD = \sqrt{\frac{\sum fx^2}{N}} = \sqrt{\frac{7978.00}{50}} = 12.63$$

2. Data from Table 3, 200 Cancellation Scores

(1) Step-Intervals Scores	(2) Midpoint X	(3) f	(4) x	(5) fx	(6) fx^2
135.5-139.5	137.5	3	18.06	54.18	978.49
131.5-135.5	133.5	5	14.06	70.30	988.42
127.5-131.5	129.5	16	10.06	160.96	1619.26
123.5-127.5	125.5	23	6.06	139.38	844.64
119.5-123.5	121.5	52	2.06	107.12	220.67
115.5-119.5	117.5	49	-1.94	-95.06	184.42
111.5-115.5	113.5	27	-5.94	-160.38	952.66
107.5-111.5	109.5	18	-9.94	-178.92	1778.46
103.5-107.5	105.5	7	-13.94	-97.58	1360.27
		$N = 200$		1063.88	8927.29

Mean = 119.44 (Table 3)

$$\frac{N}{4} = 50, \text{ hence,}$$

$$\frac{3N}{4} = 150, \text{ hence,}$$

$$Q_1 = 111.5 + \frac{1}{4} \times 4 = 115.20$$

$$Q_3 = 119.5 + \frac{3}{4} \times 4 = 123.27$$

$$Q = \frac{Q_3 - Q_1}{2} = \frac{123.27 - 115.20}{2} = 4.04$$

$$AD = \frac{\sum fx}{N} = \frac{1063.88}{200} = 5.32$$

$$SD = \sqrt{\frac{\sum fx^2}{N}} = \sqrt{\frac{8927.29}{200}} = 6.68$$

To find Q_3 we count off $\frac{3}{4}$ of N or 150 scores from the small end of the distribution. The first four steps include 101 scores, and the next step, 119.5–123.5, contains 52 scores. To fill out 150, we must take $\frac{49}{2} \times 4$, the length of the step, and add this increment (3.77) to 119.50, to locate Q_3 at 123.27. Substituting 115.20 for Q_1 and 123.27 for Q_3 in formula (6) we get a Q of 4.04.

The quartile points Q_1 and Q_3 are of importance in that they mark off the limits within which fall the *middle 50%* of the scores of the distribution. The distance between Q_1 and Q_3 is often called the *interquartile range*; and hence Q is sometimes called the *semi-interquartile range*. Q measures the average distance of the quartile points from the median, and is a valuable measure of the density with which the scores are clustered around the midpoint in the distribution. If the scores of a distribution are packed closely together the quartiles will be near to one another and Q will be small; if the scores are widely scattered, the quartiles will be relatively far apart, and Q will be large.

When the distribution is symmetrical or *normal* (p. 100), Q marks off exactly the limits of the 25% of the cases just above, and the 25% of the cases just below, the median. Accordingly, the median then lies just half way between the two quartile points Q_1 and Q_3 . In a normal distribution Q is commonly known as the *PE* (probable error). The terms Q and *PE* are often used interchangeably, but it is probably best to restrict the use of the term *PE* to the measurement of reliability (p. 113).

The steps in calculating Q when the data are grouped may be summarized as follows:

To find Q_1

- (1) Divide N by 4.
- (2) Begin at the low score end of the distribution, and count off the scores up to the interval which contains Q_1 .
- (3) Divide the number of scores necessary to locate Q_1 (i.e., to complete $N/4$) by the frequency in the interval reached in (2) above, and multiply the result by the step-interval.
- (4) Add the amount obtained in (3) to the lower limit of the step-interval within which Q_1 lies. This gives Q_1 .

To find Q_3

- (1) Find $\frac{3}{4}$ of N .
- (2) Begin at the low score * end of the distribution, and count up the scores until the interval which contains Q_3 is reached.
- (3) Divide the number of scores required to locate Q_3 by the frequency within the interval reached in (2) and multiply the result by the step-interval.
- (4) Add the amount obtained in (3) to the lower limit of the step-interval within which Q_3 lies. This gives Q_3 .

To find Q

Substitute Q_3 and Q_1 in formula (6).

3. The Average Deviation or AD

The *average deviation* or AD (also written *mean deviation* or MD and *mean variation* or MV) is the average or mean of the deviations of all the separate measures in a series taken from their central tendency (usually the arithmetic mean; less frequently the median or mode). In averaging deviations to find the AD , no account is taken of signs, and all deviations whether positive or negative are treated as positive.

An example will make our definition clearer. If we have 5 scores, 6, 8, 10, 12, and 14, the mean is easily found to be 10.

* Q_3 may also be found by counting in 25% from the high score end of the distribution. To avoid confusion, the method given above is recommended for the beginner.

It is then a simple process to find the deviation of each measure from this mean by subtracting the mean from each measure. Thus 6, the first score, minus 10 equals -4 (calculation algebraic); $8 - 10 = -2$; $10 - 10 = 0$; $12 - 10 = 2$; and $14 - 10 = 4$. The five deviations measured from the mean are $-4, -2, 0, 2$, and 4 . If we add these deviations without regard to signs the sum is 12; and dividing 12 by 5 (N), we get 2.4 as the *average* of the five deviations from their mean, or the *AD*. The formula for the *AD* when scores are ungrouped may be written

$$AD \text{ or } MD = \frac{\sum x}{N}$$

(average or mean deviation for ungrouped measures)

in which the $\sum x^*$ denotes the sum of the deviations from the mean and N is, as before, the number of cases or items.

In Table 6 the calculation of the *AD* for scores grouped into a frequency distribution is illustrated by two problems. The mean of the 50 Army Alpha scores in problem (1) has already been found in Table 3 to be 170.80. To find the *AD* of the scores in this distribution from the mean we must take our deviations (x 's) around this point. However, since the scores have been grouped into step-intervals, we are unable to get the deviation of each *separate score* from the mean. In lieu of separate score deviations, therefore, we take the deviation of the *midpoint* of each step from the mean. The substitution of the midpoint for all of the scores within a step is the only difference between the computation of x 's from grouped and from ungrouped data. The x of step 195-199, for example, is 26.20, found by subtracting 170.80 (the mean) from 197.00 (the midpoint of the step). All of the x 's are positive as far down the scale as 170-174, as in each case the midpoint is numerically larger than the mean. From the step-interval 165-169 on down to the beginning of the series, the x 's are negative, as the midpoints of these steps are all smaller than 170.80. Thus the x of step 165-169 is -3.80 ;

* The small letter x stands for the deviation of a score X from its mean (M)

and the x of the lowest step in the distribution, 140–144, is – 28.80.

It will be helpful in calculating deviations from the mean to remember that the mean is always subtracted from the individual score or midpoint value. That is, x (deviation) = X (score or midpoint) – M (mean). The calculation is algebraic. When the score or midpoint is numerically *larger* than the average the deviation is positive; when the score or midpoint is numerically *smaller* than the average the deviation is negative.

Column 4 (Table 6) gives the deviation of each step-interval, as represented by its midpoint, from the mean of the distribution. There are more scores on some steps than on others; hence each midpoint deviation in column 4 must be “weighted” or multiplied by the number of scores (f) which it represents. This gives the fx column, column 5. The first fx is 26.20; for, since there is only one score on step 195–199, we multiply the first x by 1. The next fx is 42.40, since each of the two scores on step 190–194 has an x of 21.20. In the same way we obtain the other fx 's by multiplying, in each case, the x in column 4 by its corresponding f in column 3. When all of the fx 's have been calculated the column is summed without regard to sign, and the resulting value divided by N to give the AD . In the present problem the AD equals $\frac{502.00}{50}$ or 10.04.

The formula for the AD or MD when measures are grouped into a frequency distribution is as follows:

$$AD \text{ or } MD = \frac{\sum fx}{N} \quad (8)$$

(average or mean deviation for scores grouped into a frequency distribution)

This formula applies to the calculation of the AD from the mean, median, or mode.

The second problem in Table 6 shows the calculation of the AD for 200 cancellation scores grouped into a frequency distribution in steps of 4. The mean of this distribution has been

found to be 119.44 (Table 3). Hence, the x of the topmost step, 135.5–139.5 (midpoint 137.50), from the mean is 18.06. Since the step-interval is constant, the next x may be found by subtracting 4 (the step-interval) from 18.06; and each succeeding x may be found by subtracting 4 from the x just preceding it.

The fx 's in column 5 are found, as shown in problem (1), by weighting each x by the f which it represents — by the f opposite it. The sum of the fx column is 1063.88; and, since N is equal to 200, from formula (8) we obtain 5.32 as the AD of the scores in this distribution around their mean, 119.44.

In a symmetrical or normal distribution the AD , when measured off on the scale above and below the mean, marks the limits of the middle 57.5% of the measures. The AD must always be slightly larger, therefore, than the Q which marks off the limits of the middle 50%. A large AD means that the scores of the distribution tend to scatter widely around the central tendency; a small AD that they tend to be concentrated within a relatively narrow range.

4. The Standard Deviation or SD

The *standard deviation* or SD is the most reliable measure of variability and is customarily employed in research. The SD differs from the AD in several respects. In the first place, in calculating the AD we disregard signs and treat all deviations as positive; in finding the SD , on the other hand, we avoid this difficulty of signs by squaring the separate deviations. Again, the squared deviations used in computing the SD are always taken from the mean of the distribution, and never from the median or mode. The conventional symbol used to denote the SD is the Greek letter sigma (σ).

The standard deviation or σ is the square root of the mean of the squared deviations taken from the arithmetical mean of the distribution. To illustrate the calculation of the SD in a simple case let us consider the example given on page 39 to illustrate the calculation of the AD in which the deviations of the five measures, 6, 8, 10, 12, and 14 from their mean of 10 were found

to be -4, -2, 0, 2, and 4, respectively. Squaring each of these deviations, we obtain 16, 4, 0, 4, and 16. Summing up these five squares and dividing by five, we obtain the mean of the squares (8), and, extracting the square root of this result, get 2.83, the *SD* of this series. The formula for the *SD* or σ when the series of scores is ungrouped is as follows:

$$\sigma = \sqrt{\frac{\sum x^2}{N}} \quad (9)$$

(standard deviation calculated from ungrouped data)

Table 6 illustrates the calculation of σ when scores are grouped into a frequency distribution. The process is identical with that used for ungrouped items, except that, in addition to squaring the x of each midpoint from the mean, we weight each of these squared deviations by the frequency which it represents—that is, by the frequency opposite it. This multiplication gives the fx^2 column. By simple algebra, $x \times fx = fx^2$; and accordingly the easiest way to obtain the entries in column fx^2 is to multiply the corresponding x 's and fx 's in columns (4) and (5). The first fx^2 entry, for example, is 686.44, the product of 26.20 times 26.20; the second entry is 898.88, the product of 42.40 times 21.20; and so on to the end of the column. All of the fx^2 are necessarily positive since each negative x is matched by a negative fx . The sum of the fx^2 column (7978.00) divided by N (50) gives the mean of the squared deviations as 159.56; and the square root of this result is 12.63, the *SD*. The formula for the *SD* when data are grouped into a frequency distribution is:

$$\sigma = \sqrt{\frac{\sum fx^2}{N}} \quad (10)$$

(SD or σ for data grouped into a frequency distribution)

Problem (2) of Table 6 furnishes another illustration of the calculation of σ from grouped data. In column (6), the fx^2 entries have been obtained, as in the previous problem, by multiplying each x by its corresponding fx . The sum of the fx^2

column is 8927.29; and N is 200. Hence, applying formula (10) we get 6.68 as the SD .

It was stated on page 9 that measures of variability calculated from data grouped into a frequency distribution in which the steps are five units, say, will differ slightly from the same measures calculated from the ungrouped data, or from a frequency distribution in which the same data are classified in steps of three, four, or ten units. In a frequency distribution each score within a step-interval loses its identity and is assigned the value of the midpoint of the interval. As a consequence of this, an error called the "grouping error" is introduced. When N is large and the step-interval fairly narrow this grouping error may be safely disregarded. When it is desirable to have very precise results, however, a correction — called Sheppard's correction* — may be applied to σ , provided the distribution is at least approximately normal (p 100). This correction gives a close approximation to the σ which one would obtain if the data were ungrouped. It is most useful when the step-interval is relatively wide, and the number of steps is small. The formula is

$$\sigma_{corrected} = \sqrt{\sigma_{obt}^2 - \frac{i^2}{12}} \quad (11)$$

(σ corrected for "grouping error" by Sheppard's correction)

in which σ_{obt} is the σ obtained from the frequency distribution and i is the step-interval.

Applying formula (11) to the σ 's calculated from the distributions in Table 6 we find that 12.63, the SD of the 50 Army Alpha scores, becomes $\sqrt{\frac{7978.00}{50} - \frac{25}{12}}$ or 12.55; while 6.68, the calculated SD of the 200 cancellation scores, becomes $\sqrt{\frac{8927.29}{200} - \frac{16}{12}}$ or 6.58. Sheppard's correction for grouping always reduces the calculated σ ; but generally the reduction is

* Yule, G. U., *An Introduction to the Theory of Statistics*, 9th ed., 1929, p. 212.

slight (as here) and hence is often disregarded in routine statistical calculation.

The standard deviation is less affected by sampling errors (p. 209) than is the Q and is a more stable measure of dispersion. In a normal distribution the SD , when measured off above and below the mean, marks the limits of the middle 68.26% (roughly the middle two-thirds) of the distribution. This is approximately true also for less symmetrical distributions. For example, in the first problem in Table 6 the middle 65% of the scores fall between score 183 ($170.80 + 12.63$) and score 158 ($170.80 - 12.63$).^{*} The SD is always larger than the AD which, in turn, is always larger than Q . This relationship supplies a rough check upon the accuracy of these measures of variability.

II. CALCULATION OF THE AD AND THE SD BY THE SHORT METHOD

On page 26, the Short Method of calculating the mean was outlined. This method consisted essentially in "guessing" or assuming a mean, and later applying to this value a correction which would give the actual mean. The Short Method may be used in calculating the AD and SD as well as in calculating the mean. This method is a decided time and labor saver in dealing with frequency distributions; and is well-nigh indispensable in the calculation of coefficients of correlation. The beginning student should endeavor to learn the Short Method to the point where he can use it with facility.

The Short Method of calculating the AD and SD is illustrated in Table 7. The computation of the mean by the Short Method is repeated in the table, as is also the calculation of the mean, AD , and SD by the Long Method. This will afford a readier comparison of the two procedures.

^{*} See page 132 for method of calculating the percentage of scores falling between two points in a frequency distribution.

TABLE 7

THE CALCULATION OF THE *AD* AND THE *SD* BY THE SHORT METHOD.*

DATA FROM TABLE 1. CALCULATIONS BY THE

LONG METHOD GIVEN FOR COMPARISON

1. Short Method					
(1)	(2)	(3)	(4)	(5)	(6)
Scores	Midpoint X	f	x'	fx'	fx'^2
195-199	197	1	5	5	25
190-194	192	2	4	8	32
185-189	187	4	3	12	36
180-184	182	5	2	10	20
175-179	177	8	1	8 (+ 43)	8
170-174	172	10	0		
165-169	167	6	-1	-6	6
160-164	162	4	-2	-8	16
155-159	157	4	-3	-12	36
150-154	152	2	-4	-8	32
145-149	147	3	-5	-15	75
140-144	142	1	-6	-6 (- 55)	36
		$N = 50$		98	322

$$1. AM = 172.00 \quad c = -\frac{18}{50} = -.240 \quad ci = -.240 \times 5 = -1.20$$

$$c^2 = .0576$$

$$ci_c = -1.20$$

$$M = 170.80$$

$$2. AD = \frac{\sum fx' + c(F_i - F_o)}{N} \times i \text{ (interval)}$$

$$= \frac{98 - .240(20 - 30)}{50} \times 5$$

$$= 10.04$$

$$3. SD = \sqrt{\frac{\sum fx'^2}{N} - c^2} \times i \text{ (interval)} = \sqrt{\frac{322}{50} - .0576} \times 5$$

$$= 12.63$$

* The calculation of the mean is repeated from Table 5.

TABLE 7 (continued)

2. Long Method

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Scores	Midpoint X	f	fX	x	fx	fx^2
195-199	197	1	197	26.20	26.20	686.44
190-194	192	2	384	21.20	42.40	898.88
185-189	187	4	748	16.20	64.80	1049.76
180-184	182	5	910	11.20	56.00	627.20
175-179	177	8	1416	6.20	49.60	307.52
170-174	172	10	1720	1.20	12.00	14.40
165-169	167	6	1002	-3.80	-22.80	86.64
160-164	162	4	648	-8.80	-35.20	309.76
155-159	157	4	628	-13.80	-55.20	761.76
150-154	152	2	304	-18.80	-37.60	706.88
145-149	147	3	441	-23.80	-71.40	1699.32
140-144	142	1	142	-28.80	-28.80	829.44
		$N = 50$	8540		502.00	7978.00

$$1. M = \frac{\Sigma fX}{N} = \frac{8540}{50} = 170.80$$

$$2. AD = \frac{\Sigma fx}{N} = \frac{502.00}{50} = 10.04$$

$$3. SD = \sqrt{\frac{fx^2}{N}} = \sqrt{\frac{7978.00}{50}} = 12.63$$

1. The Calculation of the AD or MD by the Short Method(1) Calculation of the AD from the Mean

The chief advantage in calculating the AD by the Short Method instead of by the Long Method lies in the fact that in the Short Method deviations are taken from an assumed mean in units of step-interval. This procedure eliminates fractions and cuts down the amount of multiplication; but at the same time it necessitates the application of a correction to the fx' and hence complicates the AD formula. The formula for finding the AD from the mean by the Short Method is

$$AD \text{ or } MD = \frac{\Sigma fx' + c(F_l - F_g)}{N} \times i \text{ (step-interval)} \quad (12)$$

(AD or MD calculated from a frequency distribution when deviations are taken from an arbitrary origin)

in which F_l equals the sum of the f 's on those steps whose mid-points are less (the subscript " l " means less) than the mean of

the distribution; F_g the sum of the f 's on those steps whose midpoints are *greater* (the subscript "g" means greater) than the mean. In Table 7 all of the midpoints from 167 down to 142, inclusive, are *less* than 170.80, the mean, and hence the F_i is 20. All of the midpoints from 172 up to 197, inclusive, are *greater* than 170.80, and the F_g is 30. It is important to remember that the F_i and the F_g are always calculated from the actual mean of the distribution — never from the assumed mean — as reference point. In consequence, the 10 scores on step 170–174, whose midpoint 172 is *greater* than 170.80, are included in the F_g . A simple check on the F_i and F_g is to make sure that $F_i + F_g = N$. Note that in Table 7, $20 + 30 = 50$.

The other terms in formula (12) require little explanation since they have all occurred before. The c is the correction in units of step-interval. It has already been found in calculating the mean and equals $-.240$. The $\Sigma fx'$ is the arithmetic sum of the fx' column and equals 98.

If now we substitute for fx' , c , F_i and F_g in formula (12), the numerator is 100.4, i.e., $[98 - .24(20 - 30)]$. Dividing this result by $50(N)$ we obtain 2.008, the AD expressed in *units of step-interval*. This value multiplied by 5 (the step length) gives 10.04, the AD of the distribution. (Compare with the AD found by the Long Method.) It is always necessary to multiply the result given in the formula proper by the step-interval, z , since both the fx' and c are in units of step-interval.

Formula (12) provides a relatively quick method of finding the AD around the mean in a frequency distribution. But the value of the formula is somewhat limited by the fact that it gives correct AD 's only when c , the step correction, is less than 1.00. In Table 7, $c = -.240$ (is less than 1.00) and the formula holds, as we find on comparing the AD 's given by the Long and Short Methods. One way of overcoming this limitation in the AD formula is to remember that no matter where the assumed mean is taken, a correction can always be applied to it to give the actual mean. If the first c calculated is less than 1.00, formula (12) may be applied directly. If, however, c is

larger than 1.00, we may assume another mean on the same step as the actual mean (now known) and take deviations from this "new" *AM*. Formula (12) will then hold

We may summarize the steps in the calculation of the *AD* from the mean by the Short Method as follows:

- (1) Find c , the correction in step-units, as shown in Table 7.
- (2) If c is less than 1.00, find the arithmetic sum of the fx' .
- (3) Calculate F_i , the total number of scores on steps with midpoints less than the mean. Next calculate F_o , the total number of scores on steps with midpoints greater than the mean.
- (4) Substitute for fx' , c , F_i , F_o , N , and i (the step length) in formula (12) to find the *AD*.

(2) Calculation of the *AD* or *MD* from the Median

It is often desirable to calculate the *AD* or *MD* from the median instead of from the mean. The same formula (12) is used in both cases. Only the method of obtaining the correction " c " to be used in the formula differs. The procedure is illustrated in Table 8 for the 200 cancellation scores taken from Table 3.

The median for the given distribution is 119.42, by the method outlined on pages 21-23. Hence we may assume or "guess" a median at the midpoint of the step-interval which contains the actual median, viz., at 117.50. Since the actual median is known, the score correction, ci , is found directly to be 1.92 by subtracting 117.50 from 119.42 (actual median — assumed median). Dividing 1.92 by 4, the step-interval, we obtain .48, the correction in *units of step*, or c .

The x' 's are taken from 117.50, the assumed median, and the fx' are obtained, as shown in Table 7, by "weighting" each x' by its corresponding f . The arithmetic sum of column (5), i.e., the fx' , is 265. F_i , the total number of scores on midpoints 117.5 to 105.5, inclusive (those less than 119.42), equals 101. And F_o , the total number of scores on midpoints 121.5 to 137.5, inclusive (those greater than 119.42), equals 99.

With fx' , c , F_i , and F_o known, the *AD* is readily found by

stituting these values in formula (12). The numerator becomes $265 + .48(101 - 99)$ or 265.96 ; and dividing by 200 and multiplying by 4 , the step length, we get 5.32 as the AD from the median.

TABLE 8

TO ILLUSTRATE THE CALCULATION OF THE AD FROM THE MEDIAN BY THE SHORT METHOD. DATA FROM TABLE 3

(1)	(2)	(3)	(4)	(5)
Scores	Midpoint \bar{X}	f	x'	fx'
15-139.5	137.5	3	5	15
15-135.5	133.5	5	4	20
15-131.5	129.5	16	3	48
15-127.5	125.5	23	2	46
15-123.5	121.5	52	1	52
15-119.5	117.5	49	0	
15-115.5	113.5	27	-1	-27
15-111.5	109.5	18	-2	-36
15-107.5	105.5	7	-3	-21
		$N = 200$		265

$$\frac{N}{2} = 100$$

Median = 119.42 (Table 3)

Assumed median = 117.50 (midpoint of step 115.5-119.5)

Correction, $ci = 119.42 - 117.50 = 1.92$

$$c = \frac{1.92}{4} = .48$$

Applying formula (12) $AD = \frac{\sum fx' + c(F_1 - F_0)}{N} \times i$ (step length)

$$AD = \frac{265 + .48(101 - 99)}{200} \times 4$$

$$= 1.33 \times 4 = 5.32$$

The Calculation of the SD by the Short Method

The calculation of the SD or σ by the Short Method is considerably easier than the calculation of the AD . The formula is

$$\sigma = \sqrt{\frac{\sum fx'^2}{N} - c^2} \times i \text{ (interval)} \quad (13)$$

(SD from a frequency distribution when deviations are taken from an assumed mean)

in which the fx'^2 is the sum of the squared deviations in units of step-interval, taken from the assumed mean, and c^2 is the squared correction in units of step-interval.

An illustration of the calculation of σ by the Short Method is given in Table 7. The first step is to fill in the fx'^2 column (column 6) by multiplying each x' in column (4) by the corresponding fx' in column (5). The process is identical with that used in the Long Method except that the x' 's are all expressed in units of step-interval. This considerably simplifies the multiplication. The calculation of c has already been described on page 28. The sum of the fx'^2 column is 322, and c^2 is .0576. Applying formula (13) we get 2.525×5 (step) or 12.63 as the σ of the distribution.

Formula (13) for the calculation of σ by the Short Method, unlike formula (12), the *AD* formula, holds good no matter what the size of c , the correction in units of step. This adds considerably to its value.

It will often save time and labor to apply the Short Method for computing σ directly to the ungrouped scores. The method is illustrated in Table 9. Note that the ten scores are ungrouped; and that it is not necessary even to arrange them in order of size. The assumed mean is taken at zero, and each score becomes at once a deviation (x') from this *AM*. The correction, c , is the difference between the actual mean (M) and the assumed mean (0), i.e. ($M - 0$); and hence c is simply the mean itself. The mean is calculated, of course, by summing the scores and dividing by N [see formula (1)]. To find σ , we square the x' 's (which are the scores), sum them, divide by N , and subtract M^2 , the correction squared. The square root of the result gives σ .

The method of calculating σ around an arbitrary mean of zero is especially useful when there are relatively few scores, say 50 or less, and when the scores are expressed in not more than two digits,* so that the squares do not become unwieldy.

* For the application of this method to the calculation of coefficients of correlation, and a scheme for reducing the size of the original scores so as to eliminate the need for handling large numbers, see page 276.

TABLE 9

TO ILLUSTRATE THE CALCULATION OF THE *SD* WHEN THE ASSUMED MEAN IS TAKEN AT ZERO, AND DATA ARE UNGROUPED

Scores	x'	$(x')^2$
18	18	324
25	25	625
21	21	441
19	19	361
27	27	729
31	31	961
22	22	484
25	25	625
28	28	784
20	20	400
<u>236</u>	<u>236</u>	<u>5734</u>

$$AM = 0$$

$$M = \frac{557.96}{23.6} = 23.6$$

$$N = 10$$

$$c = 23.6 - 0$$

$$= 23.6$$

$$c^2 = 557.96$$

$$\sigma = \sqrt{\frac{5734}{10} - (23.6)^2} \times 1 \text{ (step length)}$$

$$= \sqrt{16.44}$$

$$= 4.06$$

A calculating machine and a table of squares will greatly facilitate computation. If a calculating machine is available sum the scores as they stand and divide by N to get M . Then enter the squares of the scores in the machine in order, sum, and substitute the result in formula (13) to find σ .

III. THE COEFFICIENT OF VARIATION, V

It is often desirable to compare the variability of a given group upon two or more different tests; or to compare the variabilities of two or more groups upon the same test. We may wish, for example, to know whether 8 year old girls are more variable in height than in weight; or whether 10 year old boys are more variable than 10 year old girls in vocabulary or in memory span. The Q , AD , and SD are not suitable, ordinarily, for such comparisons. These measures give the *absolute* spread

or dispersion of test scores around their means in terms of the units of the test. But owing to differences in measuring units, we cannot compare the variability in height and the variability in weight of a given group directly; nor can we compare the relative variability in height of two groups, say boys and girls, unless the means of the two distributions are at least approximately equal. To enable us to tell whether one group is more variable than another, we need a measure which takes account *both* of the central tendency *and* of the variability of the group, and which is independent of the units in which ability is expressed. One such measure is the ratio σ/M , called the *coefficient of variation*, or V . The formula for V is

$$V = \frac{100 \times \sigma}{M} \quad (14)$$

(the coefficient of variation or coefficient of relative variability)*

The following illustrations will make the use of the formula clear. Consider, first, the case where abilities are measured in different units. A group of 7 year old boys has a mean height of 45 inches with a σ of 2.5 inches; and a mean weight of 50 lbs. with a σ of 6.0 lbs. In which trait is the group more variable, height or weight? Since we cannot compare inches and pounds directly, it is impossible to answer this question by reference to the *SD*'s of the height and weight distributions. But we can compare the relative variability of the two distributions in terms of their coefficients of variation. Thus,

$$V_{ht} = \frac{100 \times 2.5}{45} = 5.6 \quad \text{formula (14)}$$

$$\text{and} \quad V_{wt} = \frac{100 \times 6.0}{50} = 12 \quad \text{formula (14)}$$

from which it appears that these boys are 5.6/12 or 47% as variable in height as in weight.

Now let us consider the case where variability is measured in the same units, but around different points on the scale. At

* The multiplier 100 is introduced for the purpose of avoiding, as far as possible, small fractional results.

the end of 5 minutes, a group of 50 children had worked an average of 20.50 examples correctly, the σ being 5.24. At the end of 10 minutes, the same group had worked an average of 34.80 examples correctly, the σ being 9.62. If we compared the σ 's of the two distributions directly, we should probably be inclined to conclude that the group was nearly twice as variable at the end of the 10 minute period as it was at the end of the 5 minute period, since the σ has increased from 5.24 to 9.62. This conclusion is correct as far as the *absolute* spread within the group on the two occasions is concerned. But to compare the *relative* dispersion of the group in the two periods, we must take account of the fact that, with the increase in σ , the means have also increased from 20.50 to 34.80. The coefficients of variation give the following results:

$$\text{For the 5 minute period: } V = \frac{100 \times 5.24}{20.50} = 25.6$$

$$\text{For the 10 minute period: } V = \frac{100 \times 9.62}{34.80} = 27.6$$

Thus, instead of being about 50% as variable in the five minute period as in the ten, the group is $\frac{25.6}{27.6}$ or 93% as variable, when the mean score is considered as well as the absolute variability.

Objection has been raised * to the use of V in comparing the relative variability of test scores because the "true" zero point of ability in mental and educational tests is unknown. This objection does not apply, of course, to physical and physiological measures since these have true zeros. How the lack of knowledge of the true zero in a mental test may affect V can be shown most readily, perhaps, by an example. Suppose that we have given a vocabulary test to a group of children, and have obtained a mean of 25 and a σ of 5. V will equal 20. Now suppose that we add 20 very easy items, say, to our vocabulary test. It is highly probable that every child will know all of the

* Franzen, R., *Statistical Issues*, Journal of Educational Psychology, 1924, 15, pp. 367-382.

Thurstone, L. L., *The Absolute Zero in Intelligence Measurement*, Psychological Review, 1928, 35, pp. 175-197.

added words, and hence the mean score as well as every subject's score will be increased by 20 items. The absolute variability of the group (the σ) will, however, remain unchanged, as each subject occupies exactly the same relative position as before. An increase in the mean (from 25 to 45) without a corresponding increase in σ changes V from 20 to 11; and, since we could add 40 or 400 items as easily as 20, V appears to be a very unstable measure.

While theoretically correct, criticism of V because of the arbitrary nature of the zero point in mental and educational tests is not so generally destructive as it seems. Makers of standard psychological tests have been careful to begin their tests with items which, by experimental tryout, have been found to have minimal difficulty for the group for whom the test is designed. While admittedly arbitrary, such "zero" points are at least located at extremely low levels of difficulty in the ability measured by the test; hence it would be foolish to include additional easy items at the low end of the scale. The mean tells us how far the group has progressed, on the average, from the arbitrary zero point of the test. V shows, essentially, what percentage the variability is of this distance. Like M , V has a definite meaning for the test as it stands. If the range of difficulty in the test is altered, or the units changed, not only V , but M , is changed. V , therefore, is in a sense no more arbitrary than M , and the objections raised against this measure can be directed with equal force against M .

V is most useful, perhaps, in comparing the variability of a group upon the *same* test administered under different conditions; as, for example, when a group of students works at a task with and without distraction. The zero point here, at least, remains substantially constant. V may also be used to compare two or more groups on the *same* test, as when 10 year old boys and 10 year old girls are compared in tests of logical memory or picture completion. In both of these cases it is probably justifiable to assume that the "true" zero point of ability is sensibly the same for the groups compared.

It is, perhaps, most difficult to interpret V when the variability of a group upon *different* mental tests is a matter of interest. If we compare a group of girls for variability in paragraph reading and in arithmetic computation, it should be made plain that the V 's refer *only* to the specific scales upon which performance has been measured. Other tests of reading and arithmetic may — and probably will — give different results because of difference in test units, range of difficulty covered by the test, and position of arbitrary zero points. If one restricts his use of V to the particular measures which he has employed, this coefficient will furnish useful information.

IV. THE SHORT METHOD APPLIED TO DISCRETE SERIES

We have defined a truly discrete series on page 2 as one in which there are real gaps. This means that in a discrete series each measure, instead of representing an interval on a scale as in a continuous series, is a separate and distinct value. There is, for example, a real gap between one man and two men; or between one dollar and two dollars, provided the unit of measurement in the latter case is one dollar.

Table 10 illustrates the method of calculating the measures of central tendency and variability for discrete measures tabulated into a frequency distribution. The data consist of the records of the number of children in 44 families in a rural community. In the first column of the table is given the number of children in the family; in the second column — under f — the number of families of a given size. We find, for instance, one family of 10 children; three of 9; four of 8, etc. Since the measures — here, the children — are discrete, each measure must be taken at face value, and there are, in consequence, no midpoint values for the different steps. The mean is guessed at 5, and x 's are taken directly from this point. The fx' and the fx'^2 columns are calculated exactly as shown in Table 7 for continuous series — the first column is obtained by multiplying the corresponding f and x' values, and the second

by multiplying corresponding x' and fx' values. Note that since the step-interval is 1, the correction c equals cz directly.

TABLE 10

TO ILLUSTRATE THE CALCULATION OF THE MEAN, THE MEDIAN,
 Q , AD , AND SD WHEN MEASURES ARE DISCRETE

(Note that the f column gives the number of families containing the children listed in the first column)

Number of Children (Step-Intervals)	Families f	x'	fx'	fx'^2
10	1	5	5	25
9	3	4	12	48
8	4	3	12	36
7	3	2	6	12
6	5	1	5 (+ 40)	5
5	8	0		
4	7	-1	-7	7
3	4	-2	-8	16
2	4	-3	-12	36
1	2	-4	-8	32
0	3	-5	-15 (- 50)	75
	$N = 44$		90	292

$$AM = 5.00 \quad c = \frac{-10}{44} = -.23 \quad c^2 = .053$$

$$cz = -.23$$

$M = 4.77$ $N/2 = 22$; and, since the 22nd measure falls on 5, the
 $Mdn = 5$

$Mdn = 5$ $N/4 = 11$; and, since the 11th measure falls on 3, $Q_1 = 3$

$Mode = 5$ $3N/4 = 33$; and, since the 33rd measure falls between 6 and 7, $Q_3 = 6.5$

$$Q = \frac{6.5 - 3}{2} = 1.75 \quad \text{formula (6)}$$

$$AD = \frac{90 - .23(20 - 24)}{44} \times 1 \text{ (interval)} = 2.07 \quad \text{formula (12)}$$

$$SD = \sqrt{\frac{292}{44} - .053} \times 1 \text{ (interval)} = 2.57 \quad \text{formula (13)}$$

If we apply the correction $-.23$ to 5.00 (the guessed mean), 4.77, the mean of the distribution, is obtained. This result, while mathematically correct, is rather difficult to interpret in a practical way, as it is obviously impossible for a family to have four and a fraction of children. Is the median a more meaningful measure? One-half of the measures is 22, and counting in from the small end of the series we find that the

twenty-second score falls on step 5. Fractional values are, of course, really meaningless in a discrete series; and hence we simply take 5 as being roughly the median of the distribution without any interpolation. The median family, accordingly (and the modal family as well), may be said to contain 5 children, and this result on the face of it is of greater utility than the statement that the average number of children in a family is 4.77.

It is worth while examining further exactly what is meant by the statement that the mean number of children per family is 4.77. In the first place it means, of course, that the number of children in the N families examined, divided by N , gives us 4.77. But furthermore, if the families examined are actually a fair sample of *all* of the families in the "population" from which they are taken (see p. 200), it means that if we had taken *all* of these families (or another fair sample of them) the average size of the family would have been (approximately) 4.77. The mean, then, is a constant factor for the given population, such that, knowing the number of families in any fair sample of the population, we can multiply this number by the constant factor and obtain (approximately) the number of children in *all* of these families. Good use may thus be made of the mean, therefore, even when the measures are necessarily discrete — exactly the same kind of use that can be made of the mean in the case of continuous measures.

The median, on the other hand, together with the quartiles, really breaks down in the case of discrete measures. In the example above, there is actually *no* value which fulfills the definition of the median as such a point or value that one-half of the measures exceed it, and one-half fall below it. There are just 44 families in all; the median, then, would be such a point that 22 families exceeded it and 22 fell below it. Now there are 20 families falling below 5; 8 families at 5; and 16 families above 5. If we place the median exactly at 5, only 20 families instead of the required 22 fall below. And if we place the median even the least fraction above 5, the number

falling below is increased by all of the families having 5 children, so that there are then $22 + 8$ families falling below the median, or more than half. There is, in short, no median value for this series under the definition of the median which we have been using.

There is another definition of the median (see p. 20), namely, the score or measure made by the middle individual when the individuals' scores have been arranged in order from least to greatest. Strictly speaking, this definition also breaks down in the case of discrete measures since there is really no sense in speaking of two or more individuals who have the same score as being *arranged* in order of magnitude, when measures are discrete. Thus the 8 families, of 5 children each, are all exactly equal as regards number of children. Of course, we might admit that, in a sense, some one (any one) of these 8 families is the middle of the whole series, and since it is a family of 5 children, the median (so defined) is just 5, no more nor less. This is the median as we have used it. But at best it is a rough measure.

In computing the measures of variability in a discrete series, the Q is the only one which offers difficulties. In the present illustration, one-fourth $\left(\frac{N}{4}\right)$ of the measures is 11, and, counting in from the lower end of the series 11 scores, we put Q_1 on step 3 (as in the case of the median, no interpolation is made). If we check this value of Q_1 by counting in 33 scores from the upper end of the distribution, we again obtain 3 as the value of Q_1 . Three-fourths $\left(\frac{3N}{4}\right)$ of the measures is 33; and, counting in 33 scores from the lower end of the series, we complete — or count through — the frequency on step 6. If 11 scores are counted off from the other direction, we complete — or count through — the frequency on step 7. This puts Q_3 at either 6 or 7, and the best way out of the difficulty is to take Q_3 as being roughly equal to 6.5, i.e., midway between 6 and 7. Taking Q_1 equal to 3, and Q_3 equal to 6.5, Q is $\frac{6.5 - 3}{2}$ or 1.75.

The AD and σ in a discrete series are found from formulas (12) and (13) in exactly the same way as in a continuous series. For example, F_l —the number of families less than 4.77—is 20; and F_g —the number of families greater than 4.77—is 24. The AD is, therefore, $\frac{90 + (-.23)(20 - 24)}{44} \times 1$ (the step-interval) or 2.07. The σ is $\sqrt{\frac{292}{44} - .053} \times 1$ (the step-interval) or 2.57.

V. WHEN TO USE THE VARIOUS MEASURES OF VARIABILITY

1. Use the range

- (1) When the data are too scant or too scattered to justify the calculation of any other measure of variability.
- (2) When a knowledge of the total spread of the scores is all that is wanted.

2. Use the Q

- (1) For a quick, inspectional measure of variability.
- (2) When there are scattered or extreme measures
- (3) When the degree of concentration around the median itself is sought

3. Use the AD

- (1) When it is desired to weight all deviations according to their size.
- (2) When extreme deviations should influence the measure of variability, but not influence it unduly.

4. Use the SD

- (1) When the measure having the highest degree of reliability is sought *
- (2) When it is desired that extreme deviations have a proportionally greater influence upon the measure of variability.
- (3) When coefficients of correlation or measures of reliability are subsequently to be computed.†

* See discussion, pages 208-209.

† See Chapters VIII and IX.

PROBLEMS

1. Calculate the Q , AD , and σ for each of the four frequency distributions given on page 30 under problem 1, Chapter II.

2. Given the following data:

(a) $\sigma = 2.77$	$i = 1$
(b) $\sigma = 13.79$	$i = 7$
(c) $\sigma = 4.90$	$i = 5$
(d) $\sigma = 20.25$	$i = 4$

Correct each of these σ 's for "grouping error" by means of Sheppard's correction (p. 43). What two factors are of major importance in determining the size of the correction?

3. Calculate the σ of the 25 ungrouped scores given on page 14, problem 5(a), taking the AM at zero. Compare your result with the σ 's calculated from the frequency distributions of the same scores which you tabulated in step-intervals of 3 and 5 units.
4. Calculate coefficients of variation for the following traits:

Trait	Unit of measurement	Group	M	σ	V
*Length of Head	mms.	802 males	190.52	5.90	
*Body Weight	pounds	868,445 males	141.54	17.82	
Tapping Speed	M of 5 trials 30" each	68 adults, male and female	196.91	26.83	
Memory Span	No. repeated correctly	263 males	6.60	1.13	
† General Intelligence (Otis Group Intell. Scale)	Points scored	1101 adults	153.3	23.6	

Rank these traits in order for relative variability. Judged by their V 's which trait is the most variable? which the least variable? Which traits have true zeros?

* From Wechsler, D., *The Range of Human Capacities*, 1935, pp. 139-146.

† From Morton, R. L., *Laboratory Exercises in Educational Statistics*, 1928, p. 33.

5. (a) Why is the Q the best measure of variability when there are scattered or extreme scores?
 (b) Why does the σ weight extreme deviations more than does the AD ?

ANSWERS

1. (1) $Q = 3.38$ (2) $Q = 8.13$
 $AD = 3.98$ $AD = 9.01$
 $\sigma = 4.99$ $\sigma = 11.33$
- (3) $Q = 4.50$ (4) $Q = 16.41$
 $AD = 5.55$ $AD = 19.22$
 $\sigma = 7.23$ $\sigma = 24.13$
2. (a) 2.75
 (b) 13.64
 (c) 4.68
 (d) 20.22
3. σ of ungrouped scores = 6.72
 σ of scores grouped in 3-unit intervals = 6.71
 σ " " " " 5- " " = 6.78
4. V 's in order are 3.10; 12.59; 13.63; 17.12; 15.39. Ranked for relative variability from most to least: Memory Span; General Intelligence; Tapping Speed; Weight; Head Length. Last two traits have true zeros.

CHAPTER IV

GRAPHIC METHODS AND PERCENTILES

I. THE GRAPHIC REPRESENTATION OF THE FREQUENCY DISTRIBUTION

WE learned in Chapter I how scores and other measures of capacity may be organized and condensed into the tabular arrangement called a frequency distribution. We saw also how such arrangement aids in the calculation of measures of central tendency and of variability; and in general gives a better idea of the facts as a whole. Still further help in analyzing numerical data may be obtained from a graphic or pictorial treatment of our material. The advertiser has long used the graph and the illustration because he has learned that these devices catch the eye and hold the attention when the most careful array of statistical evidence fails to attract notice. For this reason also the research worker, through the medium of diagrams and graphs, attempts to utilize the attention-getting power of visual presentation; and, at the same time, to translate numerical facts — often abstract and difficult of interpretation — into more concrete and understandable form.

There are in general four methods of representing graphically measures which have been grouped into a frequency distribution. The first method gives the *frequency polygon*; the second the *histogram* or *column diagram*; the third the *cumulative frequency graph*; and the fourth the *percentile curve* or *ogive*. All of these methods will be considered in the following sections.

1. The Frequency Polygon

(1) Graphical Representation of Data; General Principles

Before considering the method of constructing a frequency polygon, let us review briefly the simple algebraic principles

which apply to all graphical representation of data. Graphing or plotting is done with reference to two lines or *coordinate axes*, the one the vertical or *Y-axis*, the other the horizontal or *X-axis*. These basic lines are perpendicular to each other, the point where they intersect being called *O*, or the *origin*. Figure 1 represents a system of coordinate axes.

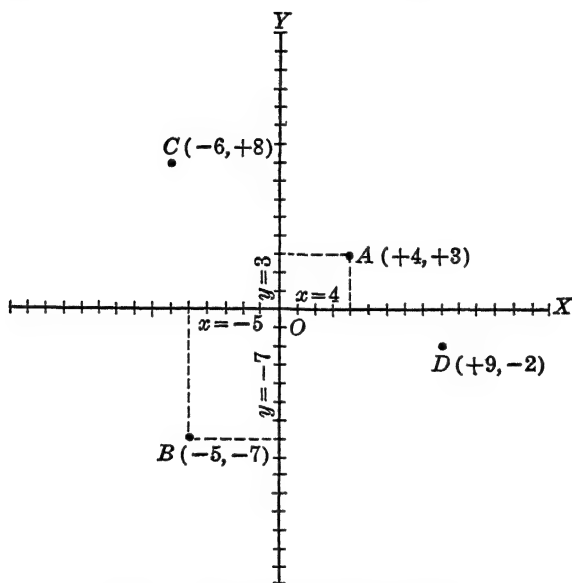


FIG. 1. A System of Coordinate Axes.

The origin is the zero point or point of reference for both axes. Distances measured along the *X-axis* to the *right* of *O* are called positive, distances measured along the *X-axis* to the *left* of *O* negative. In the same way, distances measured on the *Y-axis* *above* *O* are positive; distances *below* *O* negative. By their intersection at *O*, the *X-* and *Y-axes* form four divisions or quadrants. In the upper right division or first quadrant (see Fig. 1) both *x* and *y* measures are positive (+ +). In the upper left division or second quadrant, *x* is minus and

y plus ($- +$). In the lower left or third quadrant both x and y are negative ($- -$); while in the lower right or fourth quadrant, x is plus and y minus ($+ -$).

To locate or plot a point "A" whose coördinates are $x = 4$, and $y = 3$, we go out from O four units on the X -axis, and up from the origin three units on the Y -axis. Where the perpendiculars to these points intersect, we locate the point "A" (see Fig. 1). The point "B," whose coördinates are $x = -5$, and $y = -7$, is plotted in the third quadrant by going left from O along the X -axis five units, and then down seven units, as shown in the figure. In the same way, any point whose x and y values are known can be located with reference to OY and OX , the coordinate axes. Distances measured along the X -axis are commonly called *abscissas*; and distances along the Y -axis *ordinates*.

(2) Construction of the Frequency Polygon

Figure 2 illustrates the use of the coördinate system in the construction of a frequency polygon. This graph pictures the frequency distribution of the 50 Army Alpha scores shown in Table 1, page 5. The limits of the step-intervals (the abscissas) are laid off at regular intervals along the base line (the X -axis) from the origin; and the frequencies within each interval (the ordinates) are measured off upon the Y -axis. There is one score on the first step, 140 up to 145 (Table 1, p. 5). To represent this score on our diagram, we go out on the X -axis to 142, midway between 140 and 145, and then count up one Y -unit. The frequency on the next step-interval, 145 up to 150, is 3, hence the second point falls midway between 145 and 150, three units above the X -axis. The two scores on step 150 up to 155, the four scores on step 155 up to 160, and the frequency on each succeeding step are represented in every case by a point the specified number of scores (Y -units) above the X -axis, and midway between the upper and lower limits of the step-interval upon which the f lies. It is important in plotting a frequency polygon to remember that the midpoint of a step

is always taken to represent that particular interval. The height of the ordinate at the midpoint represents *all* of the scores within the given interval.

When all of the points have been located they are joined in regular order to give the frequency polygon shown in Figure 2. In order to complete the figure, the step next below the lowest (135 up to 140), and the step next above the highest (200 up to 205) are included on the *X*-scale. The frequency for each

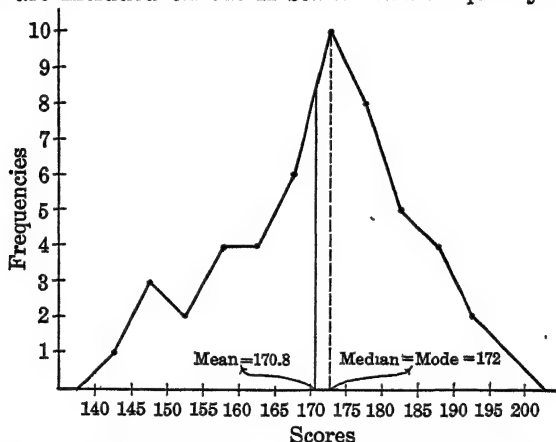


FIG 2 Frequency Polygon Plotted from the Distribution of 50 Army Alpha Scores Given in Table 1, page 5.

of these steps is zero; hence by including them we are able to begin and end the frequency polygon on the *X*-axis.

In order to give proportion and balance to his diagram, one must exercise care in the selection of unit-distances to represent the step-intervals on the *X*-axis and the frequencies on the *Y*-axis. A too-long *X*-unit tends to stretch out the polygon, while a too-short *X*-unit crowds the separate points. On the other hand, a too-long *Y*-unit exaggerates the changes from step to step, and a too-short *Y*-unit makes the polygon too flat. A good general rule is to select *X*- and *Y*-units which will make the height of the figure approximately 75% of its width. The ratio of height to width may vary from 60-80% and the figure

till have good proportions; but it can rarely go below 50% and the figure still be well balanced. The frequency polygon in Figure 2 illustrates the "75% rule." There are 13 steps or class-intervals laid off on the *X-axis* — 12 full steps plus $\frac{1}{2}$ step at the beginning and at the end of the range. Hence, our polygon should be 75% of 13, or about 10 step-interval units high. These 10 units (each equal to *one* step-interval) are laid off on the *Y-axis*. To determine how many scores (*f*'s) should be assigned to *each unit* on the *Y-axis*, we divide 10, the maximum *f* (on step 170–174) by 10, the number of intervals laid off on *Y*. The result (i.e., 1) shows that each *Y-unit* is exactly equal to one *f* or score, as shown in Figure 2. The polygon in Figure 5, page 71, furnishes another illustration of this method of plotting a frequency polygon. There are 10 intervals laid off along the base line or *X-axis* — 9 full steps plus $\frac{1}{2}$ step at the beginning and at the end of the range. The height of our figure, therefore, might be either 7 or 8 step-interval units, since 75% of 10 is 7.5. To determine the "best" value for each *Y-unit*, we first divide 52, the maximum *f* (on step 119.5–123.5) by 7, getting $7\frac{2}{7}$ or 8; and then by 8, getting 6.5 or 7. Evidently we may lay off on the *Y-axis* either 7 units, each representing 8 scores; or 8 units each representing 7 scores. The first combination was chosen because a unit of 8 *f*'s is somewhat easier to handle than one of 7; but the second combination would have given much the same figure.

The total frequency (*N*) of a distribution is represented by the *area* of its polygon; that is, by the area bounded by the frequency surface and the *X-axis*. The area of any given interval, however, cannot be taken as proportional to the number of cases within the interval because of the numerous regularities in the distribution and consequently in the frequency surface. To show the positions of the mean, median, and mode on the graph, we may locate these values on the *X-axis* and erect perpendiculars as shown in Figures 2 and 5. Note that the mode is easily located below the highest point on the frequency surface.

We may summarize the steps involved in constructing a frequency polygon as follows:

- (1) Draw two straight lines perpendicular to each other, the vertical line near the left side of the paper, the horizontal line near the bottom. Label the vertical line (the *Y-axis*) OY , and the horizontal line (the *X-axis*) OX . Put the O where the two lines intersect. This point is the *origin*.
- (2) Lay off the step-intervals of the frequency distribution at regular distances along the *X-axis*. Begin with the lower limit of the step *next below* the lowest step in the distribution, and end with the upper limit of the step *next above* the highest step in the distribution. Label the successive X distances with the step-limits. Select an X -unit which will allow all of the steps to be represented easily on the graph paper.
- (3) Mark off on the *Y-axis* successive units to represent the scores (the frequencies) on the different steps. Choose a Y -scale which will make the *maximum frequency* (the height) of the polygon approximately 75% of the width of the figure.
- (4) At the midpoint of each step-interval on the *X-axis* go up in the Y direction a distance equal to the number of scores on the step. Place points at these locations.
- (5) Join the points plotted in (4) with straight lines to give the frequency surface.

(3) Smoothing the Frequency Polygon

Because the sample is small ($N = 50$) and the distribution unsymmetrical, the frequency polygon in Figure 2 is somewhat irregular in outline. To get a better notion of how the figure would look if the data were more numerous, the frequency polygon has been "smoothed" as shown in Figure 3. There are two simple methods of smoothing a frequency surface which may be described here. In the first, a smooth curve is sketched in through as many points as possible (see Fig. 10, p. 88). In the second, a series of "moving" or "running" averages is taken and from these new or adjusted f 's or Y -values are calculated. This second method is illustrated in Figure 3. To find an adjusted or "smoothed" f , we add together the f on the given step and the f on the two adjacent steps (the one

just *below* and the one just *above*) and divide the sum by 3. For example, the smoothed f for step-interval 175-179 is $\frac{5 + 8 + 10}{3}$ or 7.67; for step-interval 155-159, $\frac{4 + 4 + 2}{3}$ or

3.33. The smoothed f 's for the other steps are given in the table below Figure 3. To find the smoothed f 's for the two steps at the extremes of the distribution, namely, 140-144, and 195-199, a slightly different procedure is necessary. Here we add together 0, the f on the step below or above, the f on the given step, and the f on the adjacent step and divide by 3. This procedure makes the smoothed f for 140-144, $\frac{0 + 1 + 3}{3}$ or 1.33, and the smoothed f for 195-199, $\frac{2 + 1 + 0}{3}$ or 1.00.

If the already smoothed f 's in Figure 3 are subjected to a second smoothing process, the outline of the frequency surface will become more nearly a continuous curve. It is doubtful, however, whether so much adjustment of the original f 's is wise. The danger in smoothing lies in the fact that when an investigator presents only a smoothed frequency polygon and not his original data, it is difficult, if not impossible, for a reader to tell with what he really started. Moreover, smoothing gives a picture of what an investigator *might* have gotten (not what he did get) if his data had been more numerous, or less subject to error than they were. If N is large and the scores or other measures reliable (p. 311) smoothing will not greatly change the shape of a graph, and hence is usually unnecessary. The frequency polygon in Figure 5, for example, which represents the frequency distribution of 200 cancellation test scores, is quite regular without any adjustment of the ordinate (i.e., the Y) values. Probably the best course for the beginning student of statistics to follow is to smooth data as little as possible. When smoothing seems to be necessary in order better to bring out the facts, one should be careful always to present original data along with the "adjusted" results.

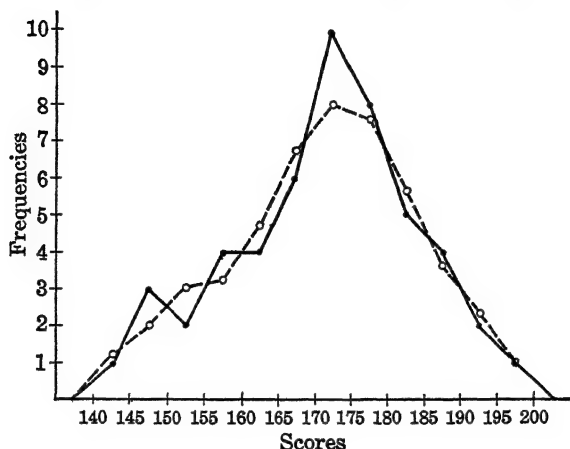


FIG 3. Original and Smoothed Frequency Polygon (Data from Table 1, p. 5) The Original and Smoothed f 's are given below the Graph.

Scores	f	Smoothed f
195-199	1	1.00
190-194	2	2.33
185-189	4	3.67
180-184	5	5.67
175-179	8	7.67
170-174	10	8.00
165-169	6	6.67
160-164	4	4.67
155-159	4	3.33
150-154	2	3.00
145-149	3	2.00
140-144	1	1.33
	<u>50</u>	

2. The Histogram or Column Diagram

A second method of representing a frequency distribution graphically is by means of a histogram or column diagram. This type of graph is illustrated in Figure 4 with the same distribution of scores represented by the frequency polygon in Figure 3. The two graphs are constructed in much the same way, with this important difference. In a frequency polygon all of the scores within a given step-interval are represented by the mid-point of that interval; while in a histogram the assumption is

made that the scores are spread uniformly over the interval and no separate concentration point is located. The measures within each interval of a histogram, therefore, are represented by a rectangle, the base of which is equal to the length of the step-interval, and the height of which is equal to the number of scores (the f) within the step-interval. Thus, the one score upon step-interval 140 to 145 is represented by a rectangle whose base equals the length of the step, and whose height

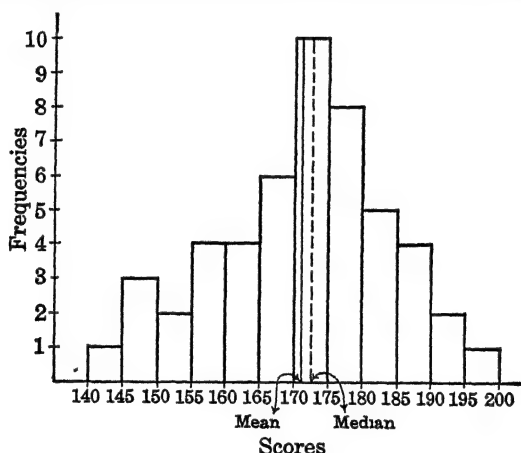


FIG. 4. Histogram of the 50 Army Alpha Scores
Shown in Table 1, page 5

equals one unit, measured off on the Y -axis. The three scores within the next interval, 145 to 150, are represented by a rectangle one interval long and three Y -units high. The altitudes of the other rectangles vary with the number of f 's upon the step-intervals, the bases being all one step long. When the same number of scores fall within two or more adjacent steps, as in the step-intervals 155 to 160, and 160 to 165, the base of the rectangle covers two or more intervals on the X -axis. The highest rectangle is, of course, that one (on step 170 to 175) which has 10, the maximum frequency, as its altitude. In selecting scales for the X - and Y -axes, the same considerations,

as to height and width of the figure, discussed on page 65 for the frequency polygon, should be observed.

Although in a histogram each step-interval is represented by a separate rectangle, it is not necessary to project the sides of the rectangles to the base line as is done in Figure 4. The rise and fall of the boundary line showing the increase or decrease in the number of scores from step to step is usually the important

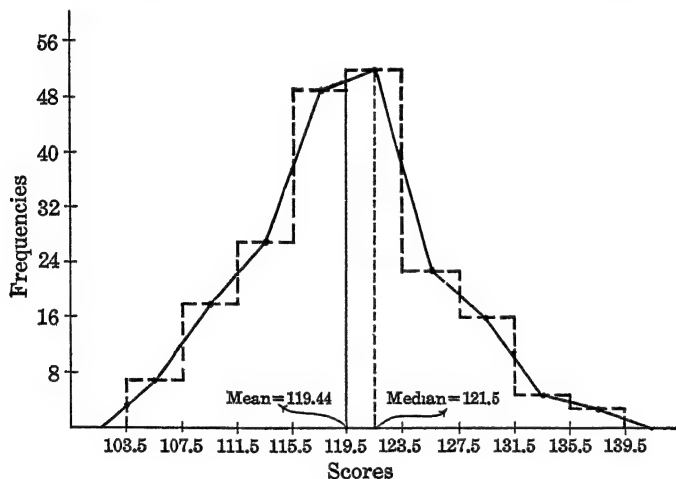


FIG 5 Frequency Polygon and Histogram of 200 Cancellation Scores Shown in Table 3, page 19.

fact to be brought out (see Fig. 5). As in a frequency polygon, the total frequency (N) is represented by the *area* of the histogram. In contrast to the frequency polygon, however, the area of *each rectangle* in a histogram is directly proportional to the number of measures within the interval. For this reason, the histogram presents a more nearly accurate picture of the relative proportions of the total frequency from step to step than does the frequency polygon.

In order to provide a more detailed comparison of the two types of frequency graph, the distribution in Table 3, page 19, is plotted upon the same coordinate axes in Figure 5 both as a

frequency polygon and as a histogram. The increased number of cases and the more symmetrical distribution of scores make these graphs more regular in appearance than the graphs in Figures 2 and 4.

The question of when to use the frequency polygon and when to use the histogram cannot be answered by giving a general rule which will cover all cases. The frequency polygon is less exact than the histogram in that it does not represent accurately i.e., in terms of area, the number of measures within the successive step-intervals. In comparing two or more frequency graphs plotted on the same diagram, however, the frequency polygon is the more useful, since the many vertical lines (and often the horizontal lines) in the two histograms will coincide. Both the histogram and the frequency polygon tell the same story and both are useful in enabling us to show in graphic fashion whether the scores of a group are distributed uniformly or whether they are piled up at the low or the high end of the scale. Not only information with regard to the group, but information with regard to the test, may be thus secured. If a test is too easy the scores will fall disproportionately at the high end of the scale; if the test is too hard the scores will pile up at the low end of the scale. If the test is neither too hard nor too easy the scores will tend to be symmetrically distributed around the mean, a few individuals scoring high, a few low, and the majority scoring somewhere near the middle of the scale. In this event, the frequency graph will approximate the "ideal" or normal frequency curve (see p. 100).

3. The Cumulative Frequency Graph

The cumulative frequency graph is another way of representing a frequency distribution by means of a diagram. Before we can plot a cumulative frequency graph, the scores of the distribution must first be added serially or cumulated, as shown in Table 11, for the two distributions taken from Table 3, page 19. These two distributions have already been used to illustrate the frequency polygon and histogram in Figures 2, 4 and 5. The

first two columns of the distributions in Table 11 are exactly the same as in any frequency distribution; but in the third column the scores have been "accumulated" successively. To illustrate, in the distribution of Army Alpha scores the first "cumulative frequency" entry is 1; $1 + 3$, from the low end of the distribution, gives 4 as the next entry; $4 + 2 = 6$; $6 + 4 = 10$, etc. The last cumulative score is, of course, equal to 50 or N , the total frequency.

TABLE 11
CUMULATIVE FREQUENCIES FOR THE TWO DISTRIBUTIONS
GIVEN IN TABLE 3, PAGE 19

Army Alpha Scores	f	Cum. f	Cancellation Scores	f	Cum. f
195-199	1	50	135.5-139.5	3	200
190-194	2	49	131.5-135.5	5	197
185-189	4	47	127.5-131.5	16	192
180-184	5	43	123.5-127.5	23	176
175-179	8	38	119.5-123.5	52	153
170-174	10	30	115.5-119.5	49	101
165-169	6	20	111.5-115.5	27	52
160-164	4	14	107.5-111.5	18	25
155-159	4	10	103.5-107.5	7	7
150-154	2	6		$N = 200$	
145-149	3	4			
140-144	1	1			
$N = 50$					

The two cumulative frequency graphs which represent the distributions of Table 11 are shown in Figures 6 and 7. Consider first the graph of the 50 Army Alpha scores in Figure 6. The step-intervals of the distribution have been laid off along the X -axis. Since there are 12 step-intervals, by the "75% rule" given on page 65, there should be 9 intervals (each equal to one step) laid off on the Y -axis. Since the maximum frequency is 50, each of these Y -units will represent $\frac{50}{9}$ or 6 scores (approximately). Instead of dividing the total Y distance into 9 intervals each representing 6 scores, however, we have, for convenience in plotting, divided our total distance into 10 intervals of 5 scores each. Note that this does not change the relation of height to width in the figure: it remains 75%.

In plotting the frequency polygon the frequency of each step was taken at the *midpoint* of the step-interval. But in constructing a cumulative frequency graph each cumulative frequency is plotted at the *upper limit* of the step upon which it falls. The first point on the curve is one Y-unit (the cumulative frequency on step 140 to 145) above 145; the second point is 4 Y-units above 150; the third 6 Y-units above 155, and so on

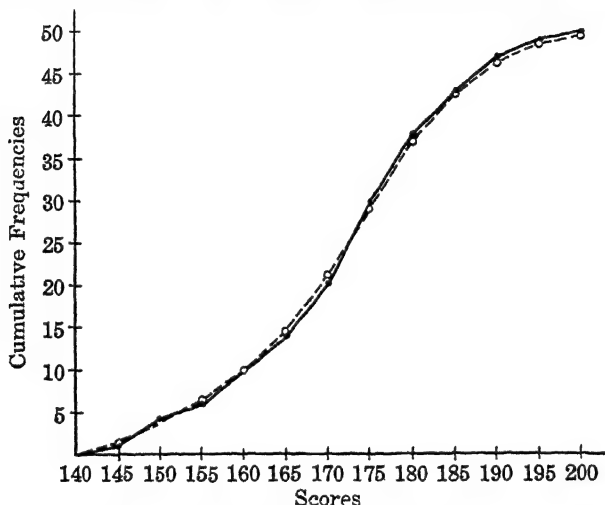


Fig. 6. Cumulative Frequency Graph (Data from Table 11, p. 73).
The dotted line represents the Smoothed Curve, the full line
the Original Data.

to the last point which is 50 Y-units above 200. The plotted points are joined in order to give the S-shaped cumulative frequency graph. In order to have the curve begin on the *X-axis* it is started at 140, the cumulative frequency of which is 0.

The cumulative frequency graph in Figure 7 has been plotted from the second distribution in Table 11 by the method described above. The curve begins at 103.5, the lower limit of the first step, and ends at 139.5, the upper limit of the last step; the cumulative frequencies, 7, 25, 52, etc., are all plotted at the *upper limits* of their respective step-intervals. The height of

this graph was determined by the "75% rule" (p. 65) as in the case of the graph in Figure 6. There are 9 step-intervals on the *X-axis*; hence, since 75% of 9 is 7 (approximately), the height of the figure is 7 step-interval units. To determine the score value of each *Y*-unit we divide 200 (maximum cumulative frequency) by 7 to give 30 (approximately). Each of the seven *Y*-units, therefore, has been taken to represent 30 scores.

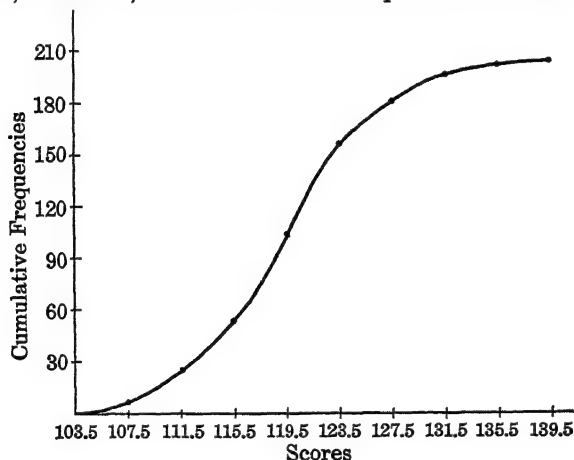


FIG. 7. Cumulative Frequency Graph
(Data from Table 11, p. 73).

The cumulative frequency graph in Figure 6 has been "smoothed" by essentially the same method as that described on page 67 for the frequency polygon. The smoothing of this graph is chiefly for illustrative purposes as it is already quite regular in appearance. The only difference between the process of smoothing a cumulative frequency graph and that of smoothing a frequency polygon is that we average *cumulative frequencies* in a cumulative frequency graph and average *adjacent frequencies* in the frequency polygon. The adjusted *Y*-value to be plotted above 150, for example, is $\frac{1 + 4 + 6}{3}$ or 3.67; above 180, $\frac{30 + 38 + 43}{3}$ or 37. The first adjusted ordinate for a cumula-

tive frequency graph is found as in smoothing a frequency polygon. Thus, the adjusted ordinate for 145 is $\frac{0 + 1 + 4}{3}$ or 1.67. To find the adjusted ordinate for 200 (highest Y -value) we must follow a slightly different procedure. Since the scores are cumulated, the f at 205 is, theoretically at least, 50 (not 0). Hence, the adjusted ordinate for 200 is $\frac{49 + 50 + 50}{3}$ or 49.67.

The cumulative frequency graph in Figure 7 has not been smoothed; it is already so regular that smoothing would be superfluous.

4. The Percentile Method and Percentile Curve

(1) The Calculation of Percentiles

We have learned (p. 21) that the median is that value in the frequency distribution below which lie 50% of the measures or scores; and that Q_1 and Q_3 mark points in the distribution below which lie, respectively, 25% and 75% of the measures or scores. In exactly the same way in which the median and quartile points were found, we may compute values below which lie 10%, 43%, or 85% of the scores. These values are called *percentiles*, and are designated, in general, by the symbol P_p , the p referring to the percentage of cases *below* the given point. Thus, P_{10} , for example, is the point below which fall 10% of the scores; P_{78} , the point below which fall 78% of the scores, etc. By analogy, it is evident that the median is P_{50} , that Q_1 is P_{25} , and that Q_3 is P_{75} .

The method of calculating a percentile value is essentially the same as that employed in calculating the median. The formula is

$$P_p = l + \left(\frac{pN - F}{f_p} \right) \times i \text{ (interval)} \quad (15)$$

(percentiles from a frequency distribution, counting from below up)
where

p = percentage of the distribution wanted, e.g., 10%, 33%, etc.

l = lower limit of the step-interval upon which P_p lies

pN = percentage of N to be counted off in order to reach P_p

F = sum of all scores upon steps below l

f_p = number of scores within the step upon which P_p falls

i = length of the step-interval

In Table 12, the nine *decile* points, P_{10} to P_{90} , have been computed by formula (15) for the distribution of scores made by the fifty college students upon Army Alpha, shown in Table 1, page 5. The details of calculation are given in Table 12. We may illustrate the method with P_{70} . Here, $pN = 35$ (70% of $50 = 35$), and from the Cum. f we find that 30 scores take us through step 170–174 up to 174.5, the *lower* limit of the step next above. Hence, P_{70} falls upon step 175–179, and, substituting $pN = 35$, $F = 30$, $f_p = 8$ (frequency upon the step), and $i = 5$ (step-interval) in formula (15), we find that $P_{70} = 177.6$ (for calculation, see Table 12). This result means that 70% of the fifty students scored *below* 177.6 upon the Army Alpha scale. The other percentile values are found in exactly the same way as P_{70} . The student should verify the calculations of the P_p in Table 12 in order to become thoroughly familiar with the method.

TABLE 12

CALCULATION OF PERCENTILES IN A FREQUENCY DISTRIBUTION

(Data are 50 Army Alpha Scores, see Table 1, p 5)

Scores	f	Cum. f	Percentiles
195–199	1	50	$P_{100} = 199.5$
190–194	2	49	
185–189	4	47	$P_{90} = 187.0$
180–184	5	43	$P_{80} = 181.5$
175–179	8	38	$P_{70} = 177.6$
170–174	10	30	$P_{60} = 174.5$
165–169	6	20	$P_{50} = 172.0$
160–164	4	14	$P_{40} = 169.5$
155–159	4	10	$P_{30} = 165.3$
150–154	2	6	$P_{20} = 159.5$
145–149	3	4	$P_{10} = 152.0$
140–144	1	1	
$N = 50$			$P_0 = 139.5$

CALCULATION OF PERCENTILES (DECILE POINTS)

10% of 50 = 5	$149.5 + \left(\frac{5 - 4}{2}\right) \times 5 = 152.0$
20% of 50 = 10	$159.5 + \left(\frac{10 - 10}{4}\right) \times 5 = 159.5$
30% of 50 = 15	$164.5 + \left(\frac{15 - 14}{6}\right) \times 5 = 165.3$
40% of 50 = 20	$169.5 + \left(\frac{20 - 20}{10}\right) \times 5 = 169.5$
50% of 50 = 25	$169.5 + \left(\frac{25 - 20}{10}\right) \times 5 = 172.0$ (<i>Mdn</i>)
60% of 50 = 30	$174.5 + \left(\frac{30 - 30}{8}\right) \times 5 = 174.5$
70% of 50 = 35	$174.5 + \left(\frac{35 - 30}{8}\right) \times 5 = 177.6$
80% of 50 = 40	$179.5 + \left(\frac{40 - 38}{5}\right) \times 5 = 181.5$
90% of 50 = 45	$184.5 + \left(\frac{45 - 43}{4}\right) \times 5 = 187.0$

It should be noted that P_0 , which marks the lower limit of the first step (namely, 139.5) lies at the beginning of the distribution. P_{100} marks the upper limit of the last step-interval, and lies at the end of the distribution. These two percentiles do not represent actual scores upon the scale, but rather limiting points. Their principal value, therefore, is to indicate the boundaries of the percentile scale.

Percentiles may be located graphically from a cumulative frequency graph, or better, perhaps, from a *percentile curve* (see later, p 84). The method of computation as applied to a smoothed cumulative frequency graph is shown in Figure 8. This graph repeats Figure 6 except for the addition of a percentile scale drawn at the right of the curve. This percentile scale was constructed by dividing the vertical line (which represents total frequency) into 100 equal parts; 0, the beginning of the scale, lies upon the *X-axis*; 100, its end, upon the curve. To find any percentile, say P_{80} , locate on the curve the point opposite 80, and directly beneath this point find P_{80} on the

X-scale. From Figure 8, P_{80} is seen to be approximately 181, which should be compared with its calculated value of 181.5 in Table 12. The median or P_{50} is readily located at 172, which is exactly its calculated value.

Any percentile, P_{17} , P_{54} , or P_{82} , may be located fairly accurately by the method outlined above. The agreement of computed and graphically determined percentiles can be im-

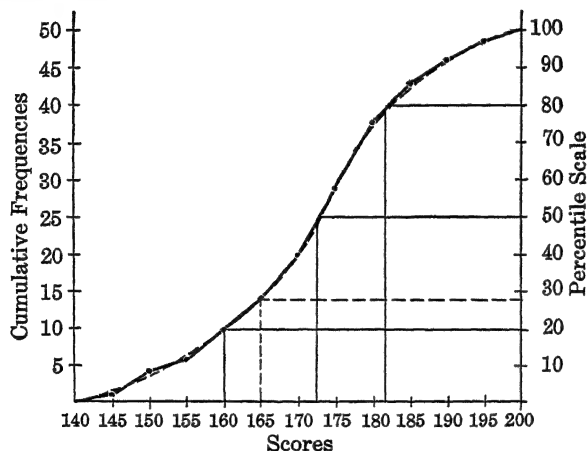


FIG. 8. To Illustrate the Method of Finding Percentiles Graphically from a Cumulative Frequency Graph.
(Data from Table 12, p 77)

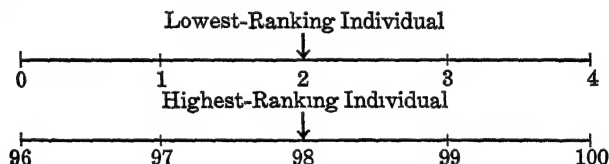
proved by drawing the diagram fairly large so that scale divisions are not crowded. Graphic methods do not yield as accurate percentile values as are obtained by calculation; but in many problems, especially when the number of scores is large, percentiles found by graphic methods are accurate enough for most purposes.

(2) The Calculation of Percentile Ranks

We have seen in the last section how percentiles, e.g., P_{15} or P_{61} , may be calculated directly from a frequency distribution or found graphically from a cumulative frequency graph. To

repeat what has been said above, these percentiles represent points in a continuous distribution below which lie given percentages of the cases. We shall now consider the problem of finding the *percentile rank* of an individual; or the position on a scale of 100 to which a subject's score entitles him. If a person's score corresponds to a percentile rank of 63, he stands sixty-third in a group of 100, or to say the same thing, he excels 63% of the members of his group.

We may first illustrate the notion of percentile rank for the simple case of scores ranked in order of merit. Suppose that a group of twenty-five persons makes scores which are ranked 1, 2, 3, 25 in order of merit. Then the highest ranking person has a percentile rank of 98, and the lowest ranking person a percentile rank of 2. How these values are calculated may be shown in the following way. On a scale running from 0 to 100, each of twenty-five individuals occupies four divisions ($100/25$ or 4%) of the scale. Hence, we assign to the poorest individual the *midpoint* of the first four divisions on the scale (0-4) or 2; to the next poorest, the midpoint of the next four divisions (4-8) or 6; and to the best person, the midpoint of the four highest divisions (96-100) or 98. Diagrams illustrating the method of assigning percentile ranks to the best and poorest persons in a group of twenty-five will make the procedure clearer:



If 100 people who have taken a test are ranked in order of merit, what is the percentile rank of the lowest ranking person? The answer is simple. Since there are just 100 subjects, each occupies one division ($100/100$ or 1%) on the percentile scale. Hence, the rank of the poorest subject is .5 (midpoint of the interval 0-1) and of the best subject 99.5 (midpoint of the

interval 99-100). These values and those of the last example may be readily found by means of the following formula* which converts scores ranked in order of merit into equivalent percentile ranks:

$$PR = 100 - \frac{(100R - 50)}{N} \quad (16)$$

*(percentile ranks for individuals ranked in order of merit
as to score)*

The R in the formula is the rank of the individual or score counting #1 as the highest rank in the group. Thus, the individual who ranks highest, i.e., #1, in a group of twenty-five has a PR of $100 - \frac{(100 \times 1 - 50)}{25}$ or 98; and the individual who ranks fifth (i.e., five from the top, twenty from the bottom) has a PR of $100 - \frac{(100 \times 5 - 50)}{25}$ or 82. The person who ranks 50th in a group of 100 has a PR of $100 - \frac{(100 \times 50 - 50)}{100}$ or 50.5, the middle of interval 50-51 on the percentile scale. Since a person's percentile rank is always the midpoint of an interval on the scale, it is evident that no one can have a percentile rank of 0 or 100. These two points are limiting values on a scale from 0 to 100.

The percentile rank of a given score in a frequency distribution may be found graphically from a cumulative frequency graph. To illustrate, suppose that in Figure 8 we wish to find the percentile rank of a student who scores 165 on Army Alpha. Locating 165 on the X -axis, we go up vertically to the curve, and from this point straight across to 29 (approximately) on the percentile scale. A student who has a score of 165, therefore, has a percentile rank of 29; he excels 29% of his group, and is excelled by 71%.

* For a table giving percentile ranks for scores ranked in order of merit, and ranging from 11 to 100 in number, see Buros, F. C., and Buros, O. K., *Expressing Educational Measures as Percentile Ranks*, Test Method Helps, #3, 1930, World Book Co., Yonkers, N.Y. In this table a rank of 1 is taken to be the highest, of 2 the next, etc.

Percentile ranks or grades (on a basis of 100) are often employed in experimental psychology when we are dealing with characters or attributes for which subjects may be ranked in order of merit, but cannot be measured directly. Thus children may be ranked for self-control, for honesty, inventiveness, and the like. These ranks when translated over into *PR*'s may be treated as scores. Percentile ranks have also been widely used in education. We are accustomed to think in terms of "percent standing." And for this reason, it is more illuminating to the teacher or parent to be told that a child has earned a *PR* of 81 upon an educational achievement battery, than to be told that his mean score is 164.3, say. A *PR* of 81 means that this pupil was excelled by only 19% of his group.

(3) The Percentile Curve or Ogive

A. Construction and Use of the Percentile Curve

The *percentile curve* or *ogive* differs from the cumulative frequency graphs in Figures 6 and 7 in two important respects. In the first place, the *X* and *Y* scales are reversed in the ogive; and in the second place, frequencies are expressed as cumulative *percents* of the distribution rather than as cumulative scores. If the percentile curve shown in Figure 9 were turned to the right through an angle of 90°, i.e., halfway around, it would become, in form at least, a cumulative frequency graph drawn from right to left.

Before plotting a percentile curve or ogive, it is necessary to change cumulative *f*'s into cumulative percents. How this is done can best be shown by an illustration. Table 13 gives the distribution of scores made upon a reading test by 125 seventh grade pupils. In column (4) of this table, the *f*'s have been cumulated from the low end of the distribution upward, as described on page 73. In column (5), each cumulative *f* has been expressed as a percentage of the total frequency, *N*(125). The conversion of Cum. *f*'s into percents may be accomplished by dividing each Cum. *f* by *N*, e.g., $2 \div 125 = .016$. But a better method — especially when a calculating

machine is available — is to determine first the reciprocal $1/N$, called the *Rate*, and then multiply each Cum. f in order by this fraction. In Table 13, the *Rate* is $\frac{1}{125}$ or .008. Hence, multiplying 2 by .008, we obtain .016; $6 \times .008 = .048$; $12 \times .008 = .096$, and so on.

TABLE 13

CALCULATION OF PERCENTILE RANKS FOR UPPER LIMITS OF
STEP-INTERVALS IN A FREQUENCY DISTRIBUTION

(The Data Represent Scores on a Reading Test Received
by 125 Seventh Grade Children)

(1)	(2)	(3)	(4)	(5)
Grades	Midpoint	f	Cum. f	Percentiles (upper limits of intervals)
74.5-79.5	77	1	125	1.000
69.5-74.5	72	3	124	.992
64.5-69.5	67	6	121	.968
59.5-64.5	62	12	115	.920
54.5-59.5	57	20	103	.824
49.5-54.5	52	36	83	.664
44.5-49.5	47	20	47	.376
39.5-44.5	42	15	27	.216
34.5-39.5	37	6	12	.096
29.5-34.5	32	4	6	.048
24.5-29.5	27	2	2	.016

$$N = 125$$

$$\text{Rate} = \frac{1}{N} = \frac{1}{125} = .008$$

CALCULATION OF Mdn , Q_1 , AND Q_3

$$Mdn = 49.5 + \frac{15.5}{36} \times 5 = 51.65$$

$$Q_1 = 44.5 + \frac{4.25}{20} \times 5 = 45.56$$

$$Q_3 = 54.5 + \frac{10.75}{20} \times 5 = 57.19$$

The graph in Figure 9 represents the ogive plotted from the data given in Table 13.* Note that the step-intervals in the

* Printed charts, called Universal Percentile Graphs, may be used conveniently in plotting percentile curves. These charts are sold by the World Book Company, Yonkers, N.Y.

graph are laid off on the *Y-axis*; and that a scale, consisting of ten equal parts each representing 10% of the distribution, is laid off on the *X-axis*. To plot the first point of the ogive, we go out 1.6% on the *X-scale* [see column (5), Table 13] and up one step-interval on the *Y-axis* to a point opposite 29.5;

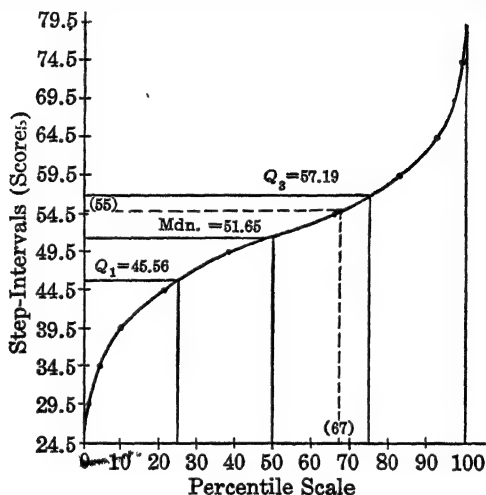


FIG. 9. Percentile Curve or Ogive Plotted from the Data in Table 13. Calculated Values of *Mdn.*, Q_1 , and Q_3 are Given for Comparison with Values Estimated from the Curve.

$$Mdn. = 51.65$$

$$Q_1 = 45.56$$

$$Q_3 = 57.19$$

to plot the second point, we go out 4.8% on the *X-scale*, and up two steps on *Y* to a point opposite 34.5. The third, and succeeding points, are plotted in the same way, until we reach 100% which lies opposite 79.5, the upper limit of the last step-interval. The outline of the graph in Figure 9 is regular enough not to require smoothing (p. 67).

The Cum. *f*'s in column (4), Table 13, tell us the number of scores in the distribution up to given points upon the scale; these points being always the *upper limits* of the steps. In the same way, the cumulative percents in column (5) tell us

the percentage of the distribution *up to* (that is, lying below) the upper limits of the succeeding step-intervals. Thus, from column (5) we know that 21.6% of our seventh graders made scores below 44.5 (upper limit of step 39.5–44.5) on the scale; that 92% made scores below 64.5 (upper limit of step 59.5–64.5) and so on. Another way of expressing the same facts is to say that 44.5 has a percentile rank of 21.6; and that 64.5 has a percentile rank of 92.

From the percentile curve or ogive, we are able to estimate (1) score values corresponding to given percentiles; and (2) percentile ranks corresponding to given scores. To obtain, for example, the median or P_{50} from Figure 9, we go up vertically from 50 on the X -scale, and then across to the Y -scale to locate P_{50} close to 52. The quartile points, $Q_1(P_{25})$ and $Q_3(P_{75})$, are found in the same way. P_{25} falls approximately at 45, and P_{75} at 57. If we compare the values of P_{25} , P_{50} , and P_{75} (namely, 45, 52, and 57) estimated from the ogive, with their computed values given in Table 13 (namely, 45.56, 51.65, and 57.19) it is clear that the agreement is quite close.

It is possible to calculate the percentile rank of a given score quite accurately from a percentile curve. To find, for example, the percentile rank of a child whose score on the reading test is 55 (see Table 13) first locate 55 on the Y -scale in Figure 9. From this point, go out to the curve, and then down vertically to the X -axis to locate the equivalent percentile rank at 67.

B. Further Applications of the Percentile Curve or Ogive

When N is large, and the step-interval not very wide, there is another method of calculating percentiles which is to be preferred to that shown in Table 13. We have learned (p. 8) that in a frequency distribution, the midpoint of a step-interval is taken to represent *all* of the scores upon that step. Because of this fact, the percentile ranks of the *midpoints* of the different step-intervals in the distribution may conveniently be taken to represent the percentile rank of *all* of the individuals whose scores fall within that step-interval. The calculation of these

"midpoint percentile ranks" is illustrated in Table 14. The first four columns repeat Table 13. The new calculations begin in column (5) wherein we determine that part of the distribution which reaches each midpoint. In column (6), these amounts are converted into midpoint percentile ranks, which represent that part of the distribution lying *below* the various midpoints * The entries in column (5) have been obtained in the following manner. There are two scores on step 24.5-29.5. The first score may be thought of as covering the first half of the step-interval, from 24.5 to 27 (midpoint); and the other score the second half of the step, i.e., from 27 to 29.5. Hence, one frequency is placed in the "Cum. f to midpoint" column opposite 27. The next entry, the Cum. f to midpoint 32, is 4. This entry is made up of the two f 's on step-interval 24.5-29.5 plus one-half of the actual f (i.e., 4) on step-interval 29.5-34.5. That is, the first two f 's take us to 29.5, the next two f 's to the midpoint of step-interval 29.5-34.5, or to 32. Each entry in column (5) is, then, the Cum f to the beginning of a step, plus *one-half* of the f on that step. Thus, Cum f to the midpoint 52 is 65. This entry is composed of the Cum f to the upper limit of the step below (i.e., 47), plus one-half of 36, the f on the step itself. All of the "midpoint-of-the-step" percentiles shown in Table 14 may be read with considerable accuracy from the ogive in Figure 9

The three children whose scores fall upon step 69.5-74.5 are all assigned the midpoint percentile rank of 98 — they excel 98% of the 125 children in the group. In like manner, the 36 children whose scores fall on step 49.5-54.5 all receive the percentile rank of 52; and all of the fifteen children whose scores fall upon step 39.5-44.5 receive the percentile rank of 15.6 or approximately 16. No child receives a percentile rank of 0 or 100. These ranks mark, respectively, the lower limit of the first or lowest step-interval, and the upper limit of the last or

* For another method of calculating midpoint percentiles, see Thurstone, L. L. *Note on the Calculation of Percentile Ranks*, Journal of Educational Psychology, 1927, 18, pp. 617-620.

TABLE 14
THE CALCULATION OF PERCENTILE RANKS FOR MIDPOINTS
OF STEP-INTERVALS IN A FREQUENCY DISTRIBUTION

(Data from Table 13, p. 83)

(1)	(2)	(3)	(4)	(5)	(6)
Grades	Midpoints	<i>f</i>	Cum. <i>f</i>	Cum. <i>f</i> to Midpt.	Midpt Percentile Ranks
74.5-79.5	77	1	125	124.5	.996
69.5-74.5	72	3	124	122.5	.980
64.5-69.5	67	6	121	118.	.944
59.5-64.5	62	12	115	109.	.872
54.5-59.5	57	20	103	93.	.744
49.5-54.5	52	36	83	65.	.520
44.5-49.5	47	20	47	37.	.296
39.5-44.5	42	15	27	19.5	.156
34.5-39.5	37	6	12	9.	.072
29.5-34.5	32	4	6	4.	.032
24.5-29.5	27	2	2	1.	.008

$$N = 125$$

$$\text{Rate} = \frac{1}{125} = .008$$

Examples:

Midpoints	Percentile Ranks
32	$4 \times .008 = .032$
52	$65 \times .008 = .520$
67	$118 \times .008 = .944$

highest step-interval. They correspond, not to score intervals upon the scale, but to limiting points. Hence, a percentile rank of either 0 or 100 is theoretically impossible.

The ogive may be employed to compare the performances of two groups in more detail than can be done from the medians and *Q*'s alone. An illustration is given in Figure 10. This graph shows the ogives representing the scores of two groups of children—a group of 200 ten year old boys and a group of 200 ten year old girls—upon an arithmetic test of 60 items. The distributions of scores and the “upper-limit-of-the-step” percentiles are tabulated below the graphs. The two ogives present an interesting comparison. The medians are rather far apart, the boys being well ahead of the girls. The boys are also

in advance of the girls in the upper half of the distribution; but at the lower end the girls are very slightly (and probably unreliably) ahead of the boys. Apparently, the brighter boys did much better than the brighter girls; while the duller boys did no better or slightly worse than the duller girls.

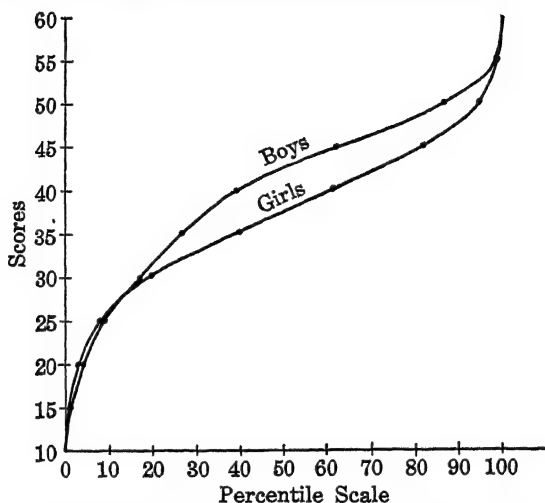


FIG. 10. Ogives Representing Comparative Scores Made by 200 ten year old Boys and 200 ten year old Girls on an Arithmetic Reasoning Test. (The curves have been slightly smoothed, see p. 67.)

Boys	<i>f</i>	Cum. <i>f</i>	Per- centiles	Girls	<i>f</i>	Cum. <i>f</i>	Per- centiles
55-59	2	200	1.000	55-59	1	200	1.000
50-54	25	198	.990	50-54	9	199	.995
45-49	48	173	.865	45-49	27	190	.950
40-44	47	125	.625	40-44	44	163	.815
35-39	26	78	.390	35-39	43	119	.595
30-34	19	52	.260	30-34	40	76	.380
25-29	15	33	.165	25-29	20	36	.180
20-24	9	18	.090	20-24	10	16	.080
15-19	7	9	.045	15-19	4	6	.030
10-14	2	2	.010	10-14	2	2	.010
$N = 200$				$N = 200$			

Another case in which the percentile curve is often useful is worth mentioning here. The scores made by a large number of

children of a given age or grade upon a standard educational examination are often represented by an ogive.* Such a curve constitutes a kind of detailed norm. The teacher or principal may evaluate the performance of a class or school against these "norms" by plotting the ogive of the scores made by his class or school upon the same graph as the "standard" ogive. Comparisons at different points and for different sections of the distribution may then be readily made.

Percentile standing or percentile rank offers a convenient method of comparing the performances of individuals upon the same test; and of comparing the performances of individuals upon tests scored in different units (p. 183). The chief disadvantage of the percentile method is the fact that units are not equal at the extremes of the scale (see p. 184 for further treatment of this topic). Over the middle of the percentile scale, however, equal changes in percentile ranking imply approximately equal changes in performance.

II. OTHER GRAPHICAL METHODS

Many problems in mental measurement, especially those which involve the determination of changes attributable to growth, practice, learning, etc., may be treated profitably by graphical methods. Two widely used graphical devices are the *line graph*, frequently employed in experimental psychology, and the *bar diagram* more often met with, perhaps, in education. These two methods will be described in this section. For a discussion of other graphical methods, the student is referred to books dealing with the subject of graphics.†

* For a detailed treatment of this topic and of percentile curves in general, see Otis, A. S., *Statistical Method in Educational Measurement*, 1925, Chaps. 5, 7 and 9.

† For a simple treatment see Rugg, H. O., *A Primer of Graphics and Statistics for Teachers*, 1925. More advanced treatments may be found in Williams, J. H., *Graphic Methods in Education*, 1924, and Karsten, K. G., *Charts and Graphs*, 1923.

1. The Line Graph

Figure 11 shows an age-progress curve. This graph represents the change in logical memory for a connected passage in

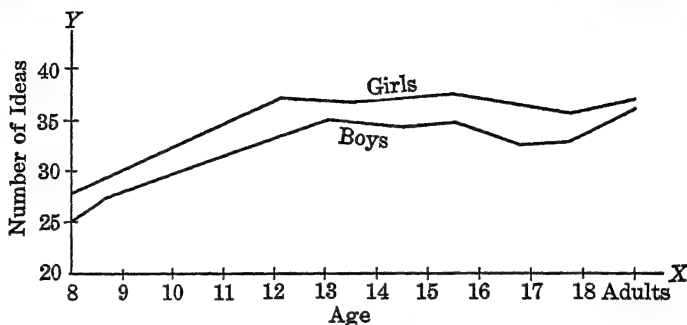


FIG. 11. Logical Memory Age is represented on *X-line* (horizontal); Score, i.e., number of ideas remembered, on *Y-line* (vertical). (After Pyle.)

boys and girls from eight to eighteen years old. Norms for adults are also included on the diagram. Age is represented on

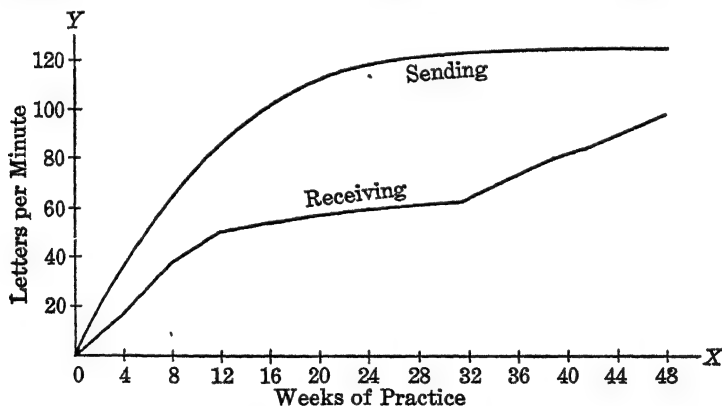


FIG. 12. Improvement in Telegraphy. Weeks of practice on *X-line*; number of letters per minute on *Y-line*. (After Bryan & Harter.)

the horizontal or *X-axis* and "average number of ideas reproduced" at each age level is marked off on the vertical or *Y-axis*.

Memory ability as measured by this test rises to a peak at year 15 for both groups after which there is a slight decline followed by a rise at the adult level. There is a small but consistent sex difference throughout, the girls being higher on the average at each age.

Figure 12 illustrates the learning or practice curve. These curves show the improvement, in sending and receiving telegraphic messages, resulting from successive trials at the same task over a period of 48 weeks. Improvement as measured by the number of letters sent or received per minute is indicated

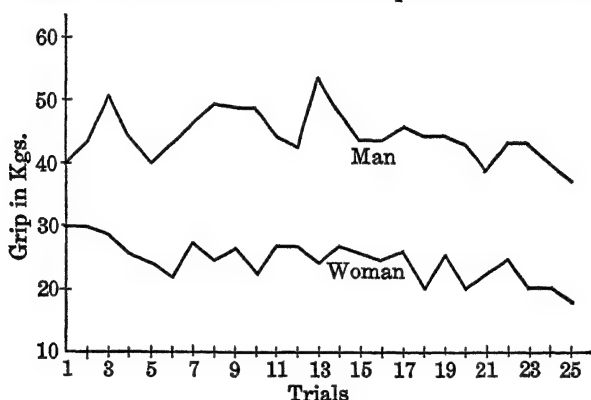


FIG. 13. Hand Dynamometer Readings in kilograms for 25 successive grips at intervals of 10 seconds. Two subjects, a man and a woman.

along the Y-axis. Weeks of practice at the given task are represented by equal intervals on the X-axis. Figure 13 is a performance or practice "curve." It represents twenty-five successive trials with the hand dynamometer made by one man and one woman. A marked sex difference in strength of grip is apparent throughout the practice period. Also a slight tendency to fatigue is evident in both subjects as the experiment progressed.

Figure 14 is Ebbinghaus' well-known "curve of retention." This curve represents memory retention as measured by the percentage of the original material retained after the passage of

different time intervals. The time intervals between learning and relearning are laid off on the *X-axis*; and the percent retained, as measured by relearning, on the *Y-axis*.

2. The Bar Diagram

The bar graph is often employed in psychology for comparing the relative amounts of some attribute (height, intelligence, educational achievement, etc.) possessed by two or more groups. In education the bar graph may be used to compare (usually

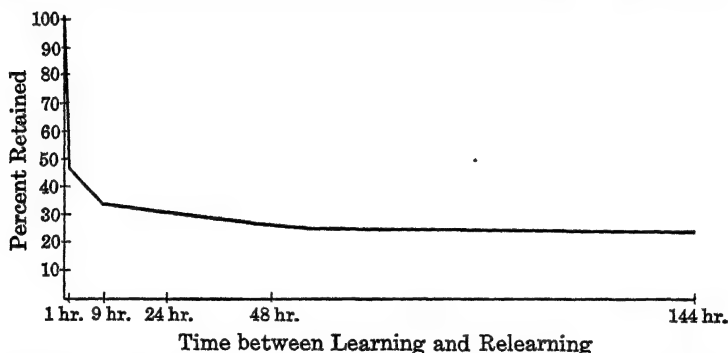


Fig. 14. Curve of Retention. The numbers on the baseline give hours elapsed from time of learning; numbers along *Y-axis* give percent retained.

in percentage terms) many different variables. Examples are: the cost of instruction in various schools or in different counties; distribution of student time in and out of school; teachers' salaries by states or districts; relative expenditures for various purposes. The commonest form of the bar graph is that in which a set of bars is used, the lengths of the bars being proportional to the amounts or percents of the quantity possessed. For emphasis, a space is usually left between the bars. The bars are drawn side by side and may be either vertical or horizontal.

A horizontal bar graph is shown in Figure 15. These bars represent the percentage of officers in various branches of the military service during the World War who received grades of *A* and *B* or *C* upon the Army Alpha Examination. The bars are

arranged in order, the group receiving the largest percent of A's and B's being placed at the top. It is clear from the diagram that the Engineers, who ranked first, received about 95% A's and B's and about 5% C's. The Veterinary Corps, which ranked last, received about 60% A's and B's and 40% C's.

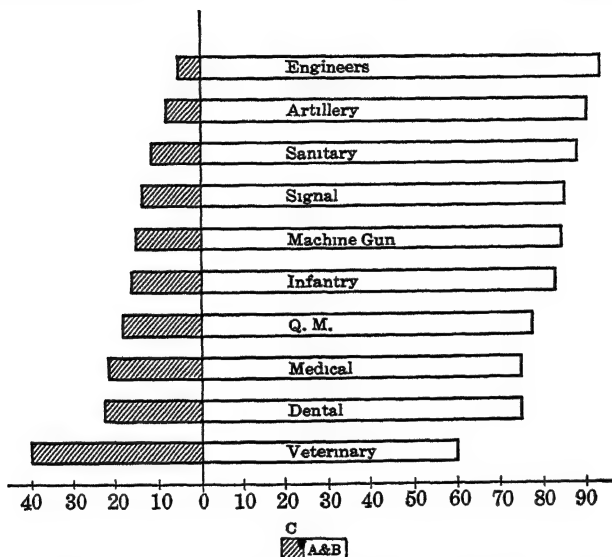


FIG. 15. Comparative Bar Graphs. The bars represent the percentage in each division of the military service receiving A's and B's or C's.

Another illustration of a bar graph is shown in Figure 16. The two parallel bars represent student enrollment in two city high schools. Each bar is divided into four parts to represent freshmen, sophomores, juniors, and seniors. The size of a division is proportional to the percentage which each class is of the whole group. This type of graph is often called a *divided-bar graph*.

School A

Freshmen	Sophomores	Juniors	Seniors
38%	31%	17%	14%

School B

Freshmen	Sophomores	Juniors	Seniors
45%	30%	16%	9%

Fig. 16. Divided Bar Graphs. The two bars represent student enrollment in two high schools. Each bar is divided into four divisions. The length of a Division shows the proportion or percentage of students in that class.

PROBLEMS

- The following distributions represent the achievement of two groups, A and B, upon a memory test.
- Plot frequency polygons for the two distributions upon the same axes.
 - Plot a histogram of Group A's scores
 - Plot cumulative frequency graphs of Group A's and of Group B's scores.

Scores	Group A	Group B
79-83	6	8
74-78	7	8
69-73	8	9
64-68	10	16
59-63	12	20
54-58	15	18
49-53	23	19
44-48	16	11
39-43	10	13
34-38	12	8
29-33	6	7
24-28	3	2
	$N = 128$	$N = 139$

Plot ogives or percentile curves for the two distributions in (1) above upon the same axes.

- (a) Determine P_{30} , P_{60} , and P_{90} graphically from each of the two ogives and compare with their calculated values.
 - (b) What is the percentile rank of score 55 in Group A's distribution? In Group B's distribution?
 - (c) A percentile rank of 70 in Group A corresponds to what percentile rank in Group B?
 - (d) What percent of Group A exceeds the median of Group B?
3. Calculate percentile ranks for the midpoints of the different step-intervals in the following distribution:

Scores	f
159.5-169.5	1
149.5-159.5	5
139.5-149.5	13
129.5-139.5	45
119.5-129.5	40
109.5-119.5	30
99.5-109.5	51
89.5-99.5	48
79.5-89.5	36
69.5-79.5	10
59.5-69.5	5
49.5-59.5	1

$$N = \overline{285}$$

4. (a) In accordance with their scores upon a learning test, 20 children are ranked in order of merit. Calculate the percentile rank of each child.
 - (b) If 60 children are ranked, what is the percentile rank of the 1st, 10th, 40th, and 60th?
5. Given the following data from five cities in the United States, represent the facts graphically by means of a bar graph.

Percent of population which is

City	Native White	Foreign-born White	Negro
A	.65	.30	.05
B	.60	.10	.30
C	.50	.45	.05
D	.40	.20	.40
E	.30	.10	.60

ANSWERS

(Calculated Values)

	Group A	Group B
2. (a) P_{30}	45.81	48.69
P_{60}	55.77	59.85
P_{90}	73.64	74.81

(d) 39% of Group A exceeds the median of Group B.

3. Midpoint percentiles in order: .9958; .9853; .9538; .8523; .7035; .5810; .4393; .2660; .1190; .0385; .0123; .0018.
4. (a) 97.5; 92.5; 87.5; 82.5; 77.5; 72.5; 67.5; 62.5; 57.5; 52.5; 47.5; 42.5; 37.5; 32.5; 27.5; 22.5; 17.5; 12.5; 7.5; 2.5.
(b) 99.17; 84.17; 34.17; .83.

Additional Problems and Questions on Chapters 1-4

1. Describe the characteristics of those distributions for which the mean is not an adequate measure of central tendency.
2. What difficulty does one encounter in calculating the median in a distribution the measures of which are discrete?
3. What is a multimodal distribution?
4. A student writes in a theme that by the application of eugenics it would be possible to raise the intelligence of the race, so that more people would be above the median *I.Q.* of 100. Comment on this statement.
5. Why cannot the σ of one test be compared directly with the σ of another test?
6. What effect will an increase in N probably have upon Q ?
7. What is the difference between a percentile and the ordinary percent grade used in school?
8. Does a percentile rank of 65 made by a given pupil mean that 65 percent of the group make scores above him; that 65 percent make the same score; or that 65 percent make scores below him?
9. What is indicated by the relatively "flat" portion of an ogive?
10. Will decreasing the number (and size) of the step-intervals used in grouping tend to make the frequency polygon more irregular?

11. Calculate the mean, median, mode, Q , AD , and SD for each of the following distributions:

(1) Scores	f	(2) Scores	f	(3) Scores	f
90-99	2	14-15	3	25	1
80-89	12	12-13	8	24	2
70-79	22	10-11	15	23	6
60-69	20	8-9	20	22	8
50-59	14	6-7	10	21	5
40-49	4	4-5	4	20	2
30-39	1			19	1
	$N = \overline{75}$		$N = \overline{60}$		$N = \overline{25}$

12. (a) Plot the distribution in 11 (1) as a frequency polygon and histogram upon the same coördinate axes.
 (b) Plot the distribution in 11 (2) as an ogive. Locate graphically the median, the Q_1 , and Q_3 .

ANSWERS

- | | | | |
|---------|----------------|-----|---------------|
| 11. (1) | Mean = 68.10 | (2) | Mean = 9.23 |
| | Median = 68.75 | | Median = 9.10 |
| | Mode = 70.05 | | Mode = 8.84 |
| | $Q = 9.01$ | | $Q = 1.69$ |
| | $AD = 10.41$ | | $AD = 2.03$ |
| | $SD = 12.50$ | | $SD = 2.48$ |
| (3) | Mean = 22.04 | | |
| | Median = 22.06 | | |
| | Mode = 22.10 | | |
| | $Q = .91$ | | |
| | $AD = 1.01$ | | |
| | $SD = 1.34$ | | |

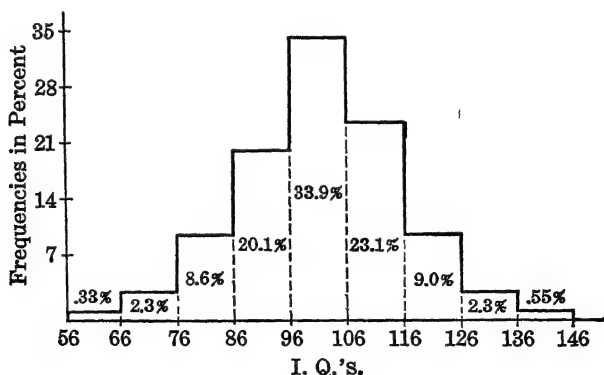
CHAPTER V

THE NORMAL PROBABILITY CURVE

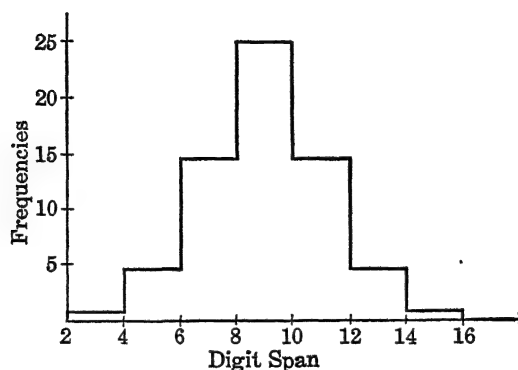
I. THE MEANING AND IMPORTANCE OF THE NORMAL DISTRIBUTION

1. Introduction

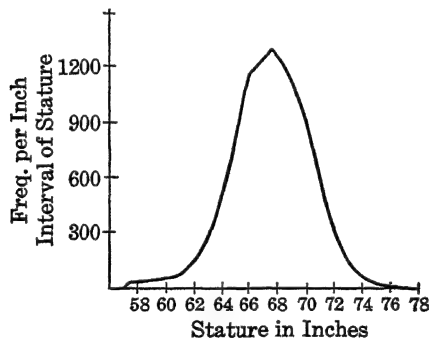
IN Figure 17 are shown four graphs, two frequency polygons and two histograms, which represent frequency distributions of data drawn from anthropometry, psychology and meteorology. It is apparent, even upon superficial examination, that all of these graphs have the same general form — the measures are concentrated closely around the center and taper off from this central high point or crest equally to the left and right. There are relatively few measures at the “low” score end of the scale; an increasing number up to a maximum at the middle position; and a progressive falling-off toward the “high” score end of the scale. If we divide the area *under* each curve (the area between the curve and the *X-axis*) by a line drawn per-



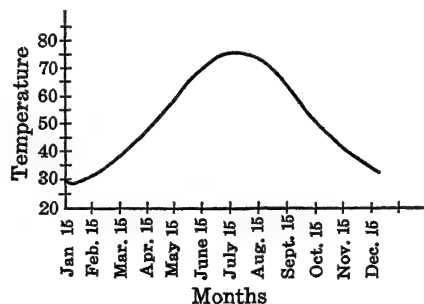
1 I Q 's of 905 unselected children, 5-14 years old (After Terman)



2 Memory span for digits, 123 adult women students (After Thorndike)



3 Statures of 8585 adult males born in British Isles (After Yule.)



4. Mean monthly temperatures, New York City, Jan-Dec, average for 47 years. (After Kelley)

FIG. 17 Frequency Distributions Drawn from Different Fields.

pendicularly through the central high point to the baseline, the two parts thus formed will be similar in shape and very nearly equal in area. It is clear, therefore, that each figure exhibits almost perfect bilateral symmetry. The perfectly symmetrical curve, or frequency surface, to which all of the figures in Figure 17 approximate, is shown in Figure 18. This bell-shaped figure is called the *normal probability curve*, or simply

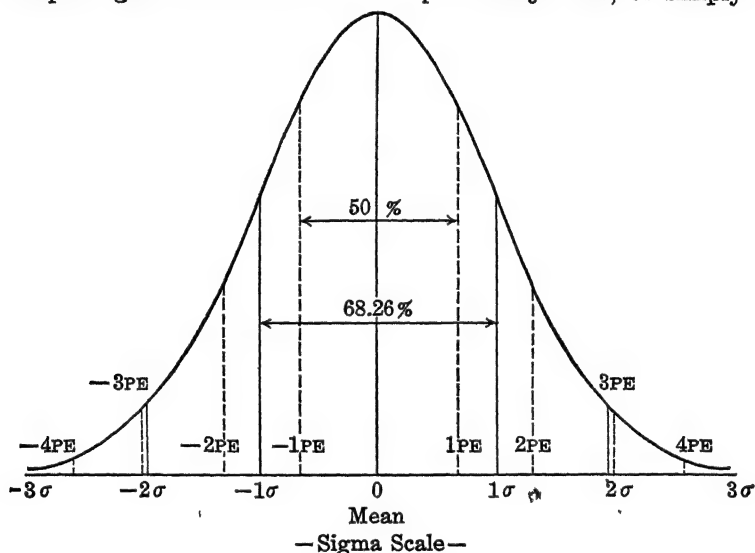


FIG 18. Normal Probability Curve.

the *normal curve*, and is of great value in mental measurement. An understanding of the characteristics of the frequency distribution represented by the normal curve is essential to the student of experimental psychology and mental measurement. This chapter, therefore, will be concerned with the "normal" distribution, and its frequency polygon, the normal probability curve.

2. Elementary Principles of Probability

Perhaps the simplest approach to an understanding of the normal probability curve is through a consideration of the

elementary principles of probability. As used in statistics, the "probability" of a particular event is defined as the expected frequency of occurrence of this event among events of a like sort. This expected frequency of occurrence may be based upon a knowledge of the conditions determining the occurrence of the phenomenon, as in dice-throwing or coin-tossing, or upon empirical data, as in mental and social measurements.

The probability of an event may be stated most simply, perhaps, as a ratio. We know, for example, that the probability of a coin falling heads is $\frac{1}{2}$, and that the probability of a die showing a two spot is $\frac{1}{6}$. These ratios are called "*probability ratios*" and are defined as that fraction the numerator of which equals the desired outcome or outcomes and the denominator of which equals the total possible outcomes. A probability ratio always falls between the limits .00 (impossibility of occurrence) and 1.00 (certainty of occurrence). Thus the probability that the sky will fall is .00; that an individual now living will some day die is 1.00. Between these limits are all possible degrees of probability which may be expressed by appropriate probability ratios.

Let us now apply this simple principle of probability to the specific case of what happens when we toss coins*. If we toss one coin, obviously it must fall either heads (H) or tails (T) 100% of the time; and furthermore, since there are only two possible outcomes, a head or a tail is *equally probable*. Expressed as a ratio, therefore, the probability of H is $\frac{1}{2}$; of T $\frac{1}{2}$; and

$$(H + T) = \frac{1}{2} + \frac{1}{2} = 1.00$$

If we toss two coins, (a) and (b), at the same time, there are four possible arrangements which the coins may take:

(1)	(2)	(3)	(4)
a b	a b	a b	a b
H H	H T	T H	T T

* Coin-tossing and dice-throwing furnish easily understood and often used illustrations of the so-called "laws of chance."

Both coins (a) and (b) may fall H; (a) may fall H and (b) T; (b) may fall H and (a) T; or both coins may fall T. Expressed as ratios, the probability of *two* heads is $\frac{1}{4}$ and the probability of *two* tails $\frac{1}{4}$. Also, the probability of an HT combination is $\frac{1}{4}$, and of a TH combination $\frac{1}{4}$. But since it makes no difference which coin falls H or which falls T, we may add these two ratios (or double the one) to obtain $\frac{1}{2}$ as the probability of an HT combination. The sum of our probability ratios is $\frac{1}{4} + \frac{1}{2} + \frac{1}{4}$ or 1.00.

Let us go a step further and increase the number of coins to three. If we toss three coins (a), (b), and (c) simultaneously there are eight possible outcomes:

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
a b c	a b c	a b c	a b c	a b c	a b c	a b c	a b c
H H H	H H T	H T H	T H H	H T T	T H T	T T H	T T T

Expressed as ratios, the probability of *three* heads is $\frac{1}{8}$ (combination 1); of *two* heads and *one* tail $\frac{3}{8}$ (combinations 2, 3, and 4); of *one* head and *two* tails $\frac{3}{8}$ (combinations 5, 6, and 7); and of *three* tails $\frac{1}{8}$ (combination 8). The sum of these probability ratios is $\frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8}$ or 1.00.

By exactly the same method used above for two and for three coins, we can figure the probability of different combinations of heads and tails when we have four, five or any number of coins. These various outcomes may be obtained in a somewhat more direct way, however, than by writing down all of the different combinations which may occur. If there are n independent factors, the probability of the presence or absence of each being the same,* the "compound" probabilities of the appearance of various combinations of factors will be expressed by the expansion of the binomial $(p + q)^n$. In this expression p equals the probability that a given event will happen, q the probability that the event will not happen, and the exponent n indicates the number of factors (e.g., coins) operating to produce the final result. If we substitute H for p

* The probability, for example, of any one of n coins showing a head (presence) or a tail (absence) is the same.

and T for q (tails = non-heads), we have for two coins $(H+T)^2$; and squaring, the binomial $(H+T)^2 = H^2 + 2HT + T^2$. This expansion may be written,

1 H^2	1 chance in 4 of 2 heads; <i>probability ratio</i>	$= \frac{1}{4}$
2 HT	2 chances in 4 of 1 head and 1 tail; <i>probability ratio</i>	$= \frac{2}{4}$
1 T^2	1 chance in 4 of 2 tails; <i>probability ratio</i>	$= \frac{1}{4}$

Total = 4

These outcomes are identical with those obtained above by listing the three different combinations possible when two coins are tossed.

If we have three independent factors operating, the expression $(p+q)^n$ becomes for three coins $(H+T)^3$. Expanding this binomial, we get $H^3 + 3H^2T + 3HT^2 + T^3$, which may be written,

1 H^3	1 chance in 8 of 3 heads; <i>probability ratio</i>	$= \frac{1}{8}$
3 H^2T	3 chances in 8 of 2 heads and 1 tail; <i>probability ratio</i>	$= \frac{3}{8}$
3 HT^2	3 chances in 8 of 1 head and 2 tails; <i>probability ratio</i>	$= \frac{3}{8}$
1 T^3	1 chance in 8 of 3 tails; <i>probability ratio</i>	$= \frac{1}{8}$

Total = 8

Again these results are identical with those got by listing the four different combinations possible when three coins are tossed.

The binomial expansion may be applied still more generally to those cases in which there are any number of independent factors operating, just so long as the probability of occurrence or non-occurrence of each factor is the same. If we toss ten coins simultaneously, for instance, we have by analogy with the above, $(p+q)^{10}$. This expression equals $(H+T)^{10}$, when H is put for the probability of a head, T for the probability of a non-head (tail), and 10 for the number of coins tossed. When the binomial $(H+T)^{10}$ is expanded,* the result is

$$H^{10} + 10H^9T + 45H^8T^2 + 120H^7T^3 + 210H^6T^4 + 252H^5T^5 \\ + 210H^4T^6 + 120H^3T^7 + 45H^2T^8 + 10HT^9 + T^{10}$$

* The student may take this expansion on faith; or he may refer to the chapter dealing with the expansion of binomials in any elementary algebra text.

which may be summarized as follows:

		<i>Probability Ratio</i>
1 H^{10}	1 chance in 1024 of all coins falling heads	$\frac{1}{1024}$
10 H^9T	10 chances in 1024 of 9 heads and 1 tail.	$\frac{10}{1024}$
45 H^8T^2	45 chances in 1024 of 8 heads and 2 tails	$\frac{45}{1024}$
120 H^7T^3	120 chances in 1024 of 7 heads and 3 tails.	$\frac{120}{1024}$
210 H^6T^4	210 chances in 1024 of 6 heads and 4 tails	$\frac{210}{1024}$
252 H^5T^5	252 chances in 1024 of 5 heads and 5 tails	$\frac{252}{1024}$
210 H^4T^6	210 chances in 1024 of 4 heads and 6 tails	$\frac{210}{1024}$
120 H^3T^7	120 chances in 1024 of 3 heads and 7 tails	$\frac{120}{1024}$
45 H^2T^8	45 chances in 1024 of 2 heads and 8 tails	$\frac{45}{1024}$
10 HT^9	10 chances in 1024 of 1 head and 9 tails	$\frac{10}{1024}$
1 T^{10}	1 chance in 1024 of all coins falling tails	$\frac{1}{1024}$

Total = 1024

These results are represented graphically in Figure 19 by a histogram and frequency polygon plotted on the same axes.

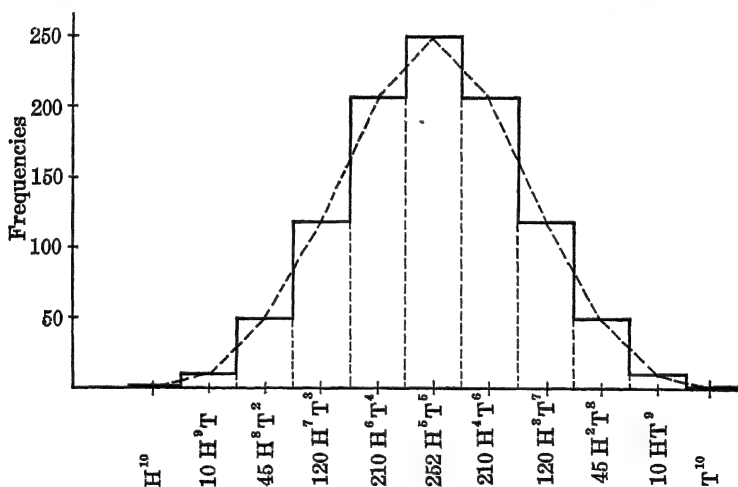


FIG. 19. Probability Surface Obtained from the Expansion of $(H + T)^{10}$.

The eleven terms of the expansion have been laid off at equal distances along the *X-axis*, and the "chances" of the occurrence of each combination of H's and T's are plotted as frequencies

on the *Y-axis*. The result is a symmetrical frequency polygon with the greatest concentration in the center and the "scores" falling away by corresponding decrements above and below the central high point. Figure 19 represents the results which we should expect *theoretically* to get by tossing ten coins 1024 times.

Many experiments have been conducted, with the idea of checking theoretical against actual results, in which coins were tossed or dice thrown a great many times. In one well-known experiment,* twelve dice were thrown 4096 times. Each 4, 5,

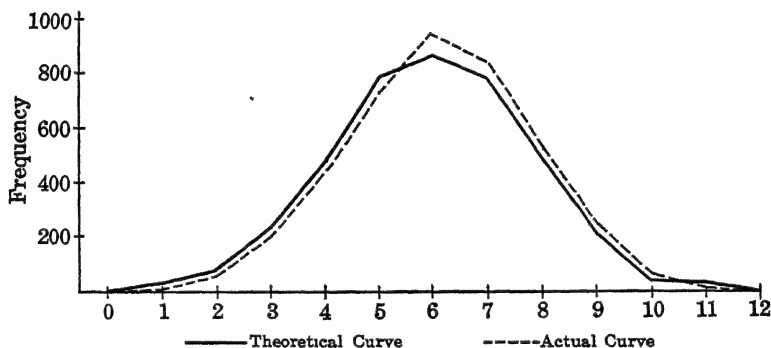


FIG. 20. Comparison of Observed and Theoretical Results in Throwing 12 Dice 4096 Times. (After Yule.)

and 6 spot was taken as a "success" and each 1, 2, and 3 spot as a "failure." Hence the probability of success and the probability of failure were the same. In a throw showing the following faces, 3, 1, 2, 6, 4, 6, 3, 4, 1, 5, 2, and 3, there would be 5 successes and 7 failures. The *observed* frequency of the different numbers of successes and the *theoretical* outcomes obtained from the expansion of the binomial expression $(p + q)^{12}$ have been plotted on the same axes in Figure 20. The student will note how closely the observed frequencies correspond to the theoretical; how near the two polygons are to being identical. If one should toss ten coins 1024 times it is probable that

* Yule, G. U., *An Introduction to the Theory of Statistics*, 9th ed., 1929, p. 258.

his results would also be in close agreement with the theoretical outcomes shown in Figure 19.

3. Why the Probability Curve is Used in Mental Measurement

The frequency curve plotted in Figure 19 from the expansion of the expression $(H + T)^{10}$ is a symmetrical ten-sided polygon. If the number of factors (e.g., coins) determining this polygon were increased from 10 to 20, to 30, and then to 40 (the baseline extent remaining the same), the sides of the polygon would also increase in number from 10 to 20, to 30, to 40. With each increase in the number of factors, the sides of the figure would become shorter, and the points on the curve would move closer together. Finally, when the number of factors became very large [when n in the expression $(p + q)^n$ became infinite] the polygon would become a perfectly smooth curve like the one in Figure 18. This "ideal" polygon or "normal" frequency curve represents the relative frequency of occurrence of various combinations of a very large number of *equal*, *similar*, and *independent* factors, when the chance of the appearance (e.g., H) or non-appearance (e.g., T) of each factor (e.g., coin) is the same.

If we compare the four graphs plotted from actual data obtained from measures of height, intelligence, memory span, and temperature, in Figure 17, with the normal probability curve in Figure 18, the similarity, as noted above, of these graphs to the normal curve is clearly evident. The resemblance of these and many other distributions to the normal curve seems to express a general tendency of quantitative data to take the symmetrical, bell-shaped form. This general tendency may be stated in the form of a "principle" as follows: measurements of natural phenomena, as well as measurements of mental and social traits, tend to be distributed symmetrically about their means in proportions which approximate those of the normal probability distribution.

Much evidence has accumulated to show that the normal distribution serves to describe the frequency of occurrence of many variable facts with a relatively high degree of accuracy.

We may classify various important phenomena which follow (at least approximately) the normal probability curve as follows:

1. *Biological statistics*: the proportion of male to female births for the same country or community over a period of years; the proportion of different types of plants and animals in cross-fertilization (the Mendelian ratios).

2. *Anthropometrical data*: height, weight, cephalic index, etc., for large groups of the same age and sex.

3. *Social and economic data*: rates of birth, marriage, or death under uniform conditions; wages and output of large numbers of workers under like conditions in the same occupation; labor costs, prices, and the like.

4. *Psychological measurements*: intelligence as measured by standard tests; speed of association, perception-span, reaction-time; educational test scores, e.g., in spelling, arithmetic, reading.

5. *Errors of observation*: measures of height, speed of movement, linear magnitudes, physical and mental traits, and the like, contain errors which are as likely to cause them to deviate above as below their true values. Chance errors of this sort follow the normal probability curve.*

The reason why many frequency distributions of scores and other measures are closely similar to those distributions obtained by tossing coins or throwing dice may be owing to the fact that the former, like the latter, are actually probability distributions. The symmetrical normal distribution, as we have seen, represents the probability of occurrence of the various possible combinations of a great many factors (e.g., coins). In a normal distribution all of the n factors are taken to be *independent, and equal in strength*; and the probability that each will be present (e.g., show H's) or absent (e.g., show T's) is the same. The appearance on a coin of a head or a tail is undoubtedly determined by a large number of small (or "chance") influences as liable to work one way as another. The twist with which the coin is tossed may be important, as well as the

* This topic is treated in Chapter VIII.

height from which it is thrown, the weight of the coin, the kind of floor upon which it falls, and many other circumstances of a like sort. By analogy, the presence or absence of each one of the probably large number of genetic factors which conceivably determines the shape of a man's head, or his intelligence, or his memory, may well depend upon a host of adventitious influences whose net effect we call "chance."

The striking similarity of obtained and probability distributions should not lead us to the conclusion that we can assume forthwith that all distributions of mental and physical traits which exhibit the normal form have *necessarily* arisen through the operation of those principles which govern the appearance of dice or coin combinations. This is an interesting speculation, to be sure, but we must be extremely cautious in making such an interpretation in advance of the facts. The factors which determine musical ability, let us say, or general mechanical skill are too little known to warrant the assumption, *a priori*, that they combine in the same proportions as do the head and tail combinations in "chance" distributions of coins. The selection of the normal curve rather than some other type of curve is, after all, sufficiently justified by the fact that this curve generally does fit the data better. But the "theoretical justification and the empirical use of the normal curve are two quite different matters." *

II. TABLES OF FREQUENCIES OF THE NORMAL PROBABILITY DISTRIBUTION

The equation of the normal probability curve may be written as

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \quad (17)$$

(equation of the normal probability curve)

in which x represents the scores (expressed in the form of deviations from the mean) laid off along the baseline or X -axis;

* Jones, D. C., *A First Course in Statistics*, 1921, p. 233.

and y (the height of the curve above the baseline) represents the frequency of a given x -value, or the number of individuals achieving a given score (see Fig. 24, p. 125). The other terms in the equation are *constants*: e , which equals 2.7183, is the base of the Napierian system of logarithms; π (3.1416) is the ratio of the circumference of a circle to its diameter; N is the number of cases; and σ is the standard deviation of the distribution.

It is possible from equation (17) to compute the frequency (y) of a given value x , i.e., the number of individuals making a certain score; and also the number, or percentage, of individuals scoring between two points, or above or below a given point in the distribution. But this is not an especially easy task, and fortunately is rarely necessary, as tables are available from which the frequency of any fractional part of a normal distribution may be readily obtained. A knowledge of these tables (Tables 15 and 16) is extremely valuable in the solution of a large number of problems. For this reason, before we consider any problems in which normality of distribution in the data treated may be reasonably assumed, it is very desirable that the construction and use of Tables 15 and 16 be clearly understood.

1. The Construction and Use of Tables 15 and 16

Table 15 gives the fractional parts of the total area under the normal curve found between the mean and ordinates (y 's) erected at various distances from the mean. In Table 15 these distances are measured in σ units.* The total area of the curve in Table 15 (the number of cases in the distribution) is taken arbitrarily to be 10,000, because of the greater ease with which fractional parts of the total area may then be calculated.

The first column of the table, $\frac{x}{\sigma}$, gives the distances in tenths of σ measured off on the baseline of the normal curve from the mean as origin; distances in hundredths of σ are given

* Table 15 should be studied in conjunction with Figure 18.

TABLE 16

FRACTIONAL PARTS OF THE TOTAL AREA (TAKEN AS 10,000) UNDER THE NORMAL PROBABILITY CURVE, CORRESPONDING TO DISTANCES ON THE BASELINE BETWEEN THE MEAN AND SUCCESSIVE POINTS LAID OFF FROM THE MEAN IN UNITS OF PE

Example: we find between the mean and a point $1.55 PE$ ($\frac{x}{PE} = 1.55$) from the mean 35.21% of the entire area under the curve.

$\frac{x}{PE}$	00	.05	$\frac{x}{PE}$.00	.05
0	0000	0135	3.0	4785	4802
.1	0269	0403	3.1	4817	4832
.2	0537	0670	3.2	4846	4858
.3	0802	0933	3.3	4870	4881
.4	1063	1193	3.4	4891	4900
.5	1320	1447	3.5	4909	4917
.6	1571	1695	3.6	4924	4931
.7	1816	1935	3.7	4937	4943
.8	2053	2168	3.8	4948	4953
.9	2281	2392	3.9	4957	4961
1.0	2500	2606	4.0	4965	4968
1.1	2709	2810	4.1	4972	4974
1.2	2909	3004	4.2	4977	4979
1.3	3097	3187	4.3	4981	4983
1.4	3275	3360	4.4	4985	4987
1.5	3442	3521	4.5	4988	4989
1.6	3597	3671	4.6	4990	4991
1.7	3742	3811	4.7	4992	4993
1.8	3876	3939	4.8	4994	4995
1.9	4000	4058	4.9	4995	4996
2.0	4113	4166	5.0	4996	4997
2.1	4217	4265	5.1	4997.1	4997.4
2.2	4311	4354	5.2	4997.7	4998
2.3	4396	4435	5.3	4998.2	4998.5
2.4	4473	4508	5.4	4998.6	4998.8
2.5	4541	4573	5.5	4999	4999.1
2.6	4603	4631	5.6	4999.2	4999.3
2.7	4657	4682	5.7	4999.4	4999.5
2.8	4705	4727	5.8	4999.54	4999.6
2.9	4748	4767	5.9	4999.65	4999.7

by the headings of the columns. To find the number of cases in a normal distribution between the mean and the ordinate erected at a distance of 1σ from the mean, we go down the x column until 1.0 is reached, and in the next column under .00 take the entry opposite 1.0, viz., 3413. This figure means that there are 3413 cases in 10,000, or 34.13% of the entire area of

the curve, between the mean and 1σ . Put more exactly, 34.13% of the cases in a normal distribution fall within the interval bounded by the baseline of the curve, the ordinate erected at the mean, the ordinate erected at a distance of 1σ from the mean, and the curve itself (see Fig. 18 for illustration). To find the percentage of the distribution between the mean and 1.57σ , say, we go down the $\frac{x}{\sigma}$ column to 1.5, then across horizontally to the column headed .07, and take the entry 4418. This means that in a normal distribution, 44.18% of the entire distribution fall between the mean and 1.57σ .

We have so far considered only σ distances measured in the *positive* direction from the mean; that is, we have taken account only of the *right* half — the high score end — of the normal curve. Since the curve is bilaterally symmetrical, the entries in Table 15 apply also to σ distances measured in the *negative* (to the *left*) as well as in the positive direction. Accordingly, to find the percentage of the distribution between the mean and -1.26σ we take the entry in the column headed .06, opposite 1.2 in the $\frac{x}{\sigma}$ column. This entry (3962) tells us that 39.62% of the cases in the normal distribution fall between the mean and -1.26σ . The percentage of cases between the mean and -1σ is 34.13; and the student will now be able to verify the statement made on page 44 that between the mean and $\pm 1\sigma$ are found 68.26% of the cases in a normal distribution (see also Fig. 18 for illustration).

While theoretically the normal curve meets the baseline at infinite distances to the right and left of the mean, for practical purposes the curve may be taken to end at points -3σ and $+3\sigma$ distant from the mean. Table 15 shows that 4986.5 cases in the total 10,000 fall between the mean and $+3\sigma$; and 4986.5 cases will, of course, fall between the mean and -3σ . Therefore, 9973 cases in 10,000, or 99.73% of the entire distribution, fall within the limits set by -3σ and $+3\sigma$.

By cutting off the curve at these two points, therefore, we disregard only .27 of 1% of the distribution, a negligible amount except in very large samples.

Instead of σ the PE may be used as the unit of measurement in determining the theoretical frequency within given intervals of the normal curve. Table 16 gives the fractional parts of the total area under the normal curve found between the mean and ordinates erected at various PE distances from the mean. This table is read in exactly the same way as Table 15. To find, for instance, the number of cases between the mean and $1PE$ (or more accurately the ordinate erected at this point), we go

down the $\frac{x}{PE}$ column to 1.0 and opposite this entry in the next column headed .00 read 2500. Twenty-five percent of the cases in the distribution, therefore, lie between the mean and $1PE$. In like manner, 25% of the cases lie between the mean and $-1PE$; hence it is clear that the middle 50% of a normal distribution fall between the mean and $-1PE$ and $+1PE$ measured off from the mean (p. 37). Table 16 does not read in as fine units as Table 15, only tenths and .05ths PE divisions being given. If smaller divisions are desired, interpolation can be readily made.

Just as it is customary to disregard that part of a normal curve beyond the limits $\pm 3\sigma$, we ordinarily ignore also that part of the normal curve beyond the limits $\pm 4PE$. There are 9930 (4965×2) cases in the total 10,000 between the mean and $\pm 4PE$ (Table 16). Hence, in cutting off the curve at $\pm 4PE$, we disregard only .70 of 1% of the cases in the entire distribution.

There is little to choose as between Tables 15 and 16. Table 15 admits of easy interpolation and is more often used in mental measurement; but Table 16 is probably accurate enough, without interpolation, for most purposes.

2. Some Important Properties of the Normal Probability Curve

In the normal probability curve, the mean, the median, and the mode all fall exactly at the midpoint of the distribution

and are numerically equal. Since the normal curve is bilaterally symmetrical, all of the measures of central tendency must coincide at the middle of the distribution.

The measures of variability include certain constant fractions of the total area of the normal curve, which may be read from Tables 15 and 16. Between the mean and $\pm 1\sigma$ lie the middle two-thirds (approximately) of the cases in the normal distribution. Between the mean and $\pm 2\sigma$ is found 95% (approximately) of the distribution; and between the mean and $\pm 3\sigma$ is found 99.7% (approximately 100%) of the distribution. There are 68 chances (approximately) in a hundred that a score will lie within $\pm 1\sigma$ from the mean in the normal distribution; there are 95 chances in 100 that it will lie within $\pm 2\sigma$ from the mean; and 99.7 chances in a hundred that it will lie within $\pm 3\sigma$ from the mean.

As we have seen, $\pm 1PE$ mark off the middle 50% of the cases, i.e., the 25% of the measures directly above, and the 25% directly below, the measure of central tendency. Furthermore, $\pm 2PE$ include 82.26% of the measures in the distribution; $\pm 3PE$, 95.70% of the measures in the distribution; and $\pm 4PE$, 99.30% of the measures in the distribution.

The following constant relations exist among the measures of variability:

1. $PE = .6745\sigma$
2. $\sigma = 1.4826PE$
3. $PE = .8453AD$
4. $\sigma = 1.2533AD$
5. $AD = .7979\sigma$
6. $AD = 1.1843PE$

These equations may be verified (for σ and PE) * from the percents of area included by each. Thus, we find by interpolation in Table 15 that $.6745\sigma$ ($1PE$) includes the 25% of the distribution just above (or below) the mean.

* A table giving percentages of area between the mean and various points in the distribution in terms of AD is not given, as the AD is infrequently used as compared with the SD or PE . From equation (5) above, it may be found, however, that 57.5% of the cases in a normal distribution fall between the mean and $\pm 1AD$.

The relations among the measures of variability, given above, are arranged in order of importance; the first is probably the only one employed often enough to warrant its being memorized. From these formulas it is evident why it was stated earlier (p. 44) that σ is greater than AD which in turn is greater than $Q(PE)$.

III. THE MEASUREMENT OF DIVERGENCE FROM NORMALITY

.. Skewness

In a frequency polygon or histogram the first thing which strikes the eye is the symmetry or, what is more often the case, the lack of symmetry in the figure. As pointed out above, in the normal curve the mean, the median, and the mode all coincide and there is perfect balance between the right and left halves of the figure. A distribution is said to be "skewed" when the mean, the median and the mode fall at different points in the distribution, and the balance (or center of gravity) is shifted to one side or the other, to right or left. It is often important to know (1) whether the skewness which so often occurs in distributions of test scores and other measures represents a real divergence from the ideal normal curve; or (2) whether it is the result of chance fluctuation, arising from temporary causes, and is not significant of real disagreement. The degree of displacement or skewness in a frequency distribution may be measured by the formula

$$Sk = \frac{3(\text{mean} - \text{median})}{\sigma} \quad (18)$$

(a measure of skewness in a frequency distribution)

In a normal distribution the mean equals the median and the skewness is 0. The more nearly the distribution approaches the normal form, the closer together are the mean and the median, and the less the skewness. Distributions are said to be skewed *negatively*, or to the *left*, when the scores are massed at the high end of the scale (the right end), and spread out

gradually at the low or left end, as shown in Figure 21. Distributions are skewed *positively*, or to the *right*, when the scores are massed at the low (the left) end of the scale, and spread out gradually toward the high or right end as shown in Figure 22.

If we apply formula (18) to the distribution of 50 Army Alpha scores in Table 1, page 5, $-.28$ is obtained as a measure of skewness. This result points to a slight negative skewness in the data, which may best be seen by reference to Figure 2, page 65. Formula (18) gives a measure of skewness for the

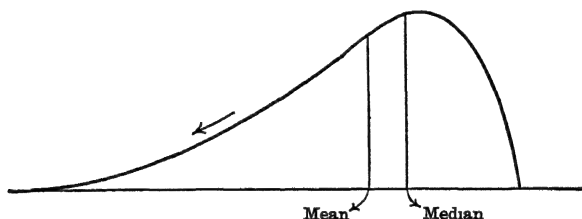


FIG 21. Negative Skewness: To the Left.

distribution of the 200 cancellation scores (Table 3, p. 19) as $.009$. This very low degree of positive skewness shows how closely this distribution approaches the symmetrical probability form.

There is another measure of skewness given by the formula

$$Sk = \frac{(P_{90} + P_{10})}{2} - P_{50} \quad (19)$$

(a measure of skewness in terms of percentiles)*

Applying this formula to the distributions of 50 Army Alpha scores and 200 cancellation scores, we obtain for the first distribution a $Sk = -2.50$; and for the second distribution a $Sk = .03$. These results are numerically different from the measures of skewness obtained from formula (18), because

* Kelley, T. L., *Statistical Method*, 1923, p. 77. The terms in this formula, as given by Kelley, have been reversed so that the sign of Sk will agree with the conventional notion of positive and negative skewness.

the two measures of skewness are computed from different reference values in the distribution, and hence are not directly comparable. The two formulas agree, however, in indicating slight negative skewness for the distribution of 50 Army Alpha scores, and a negligible degree of positive skewness for the 200 cancellation scores. In comparing the skewness of two distributions we should use either formula (18) or (19); not first the one and then the other.

The important question of how much skewness a distribution must exhibit before it may be said to be *significantly*

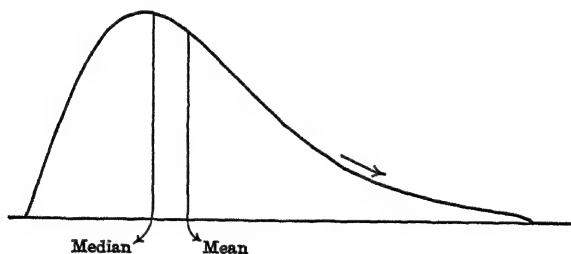


FIG. 22. Positive Skewness: To the Right

skewed can be answered only when we have calculated a "standard error" of our measure of skewness. A formula for the standard error of Sk , as given by formula (19), and a method of testing whether the skewness of a given distribution is significant or not is discussed in Chapter VIII, page 229.

2. Kurtosis

The term kurtosis refers to the "peakedness" or flatness of a frequency distribution as compared with the normal. A frequency distribution more peaked than the normal is said to be *leptokurtic*; one flatter than the normal, *platykurtic*. Figure 23 shows a leptokurtic distribution and a platykurtic distribution plotted on the same diagram around the same mean. A normal curve (called *mesokurtic*) has also been drawn in on the diagram

to bring out the contrast in the figures, and to make comparison easier. A formula for measuring kurtosis is

$$Ku = \frac{Q}{(P_{90} - P_{10})} \quad (20)$$

(a measure of kurtosis in terms of percentiles)

For the normal curve, formula (20) gives $Ku = .26315$. If Ku is greater than .26315 the distribution is platykurtic; if Ku

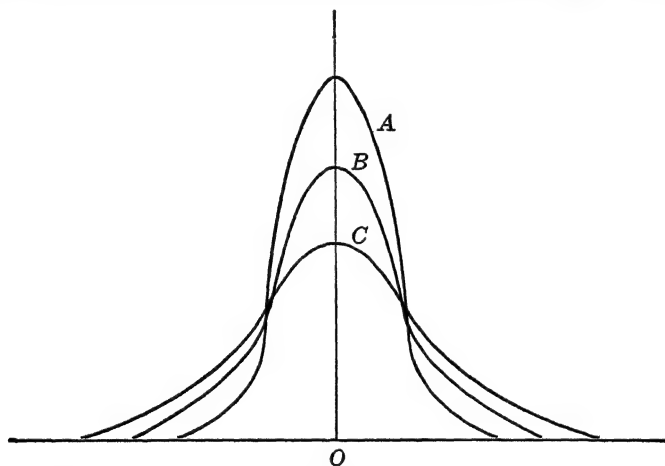


FIG. 23. Leptokurtic (A), Mesokurtic or Normal (B), and Platykurtic (C) Curves.

is less than .26315 the distribution is leptokurtic. Calculating the kurtosis of the distributions of 50 Alpha scores and 200 cancellation scores, discussed above, we obtain $Ku = .237$ for the first distribution, and $Ku = .223$ for the second. Both distributions, therefore, are slightly leptokurtic. To determine whether the kurtosis in a distribution is significant, that is, whether the curve is too high or too flat to be treated as sensibly normal, we must evaluate Ku in terms of its standard error. A formula for the standard error of Ku , and a method of determining the significance of an obtained Ku are given in Chapter VIII, page 230.

3. Testing a Given Frequency Distribution for Divergence from Normality

If an investigator finds that his distributions are not significantly skewed, nor significantly flattened or peaked, he is usually justified in treating his data as "normally distributed" as far as most problems are concerned. In research wherein great accuracy is desired, however, more exact methods of testing the normality of a distribution are demanded. The usual procedure in making an exact test is to find that normal curve — equivalent in area (N) to the given distribution and having the same σ and the same mean — which "best fits" the given data. The frequencies in the two distributions, normal and obtained, are then compared to discover whether the differences between them are too large to be attributed to fluctuations of sampling (see p. 210), and are therefore significant. A comparison "by eye" of the actual and theoretical figures can be made by plotting the two distributions on the same axes in the manner described in the next section. But unless the differences between the curves are very small or very large, mere inspection of this sort is not a safe nor an exact method to use in judging the significance of the difference between two distributions.

One precise method of comparing an obtained distribution with the normal distribution fitted to the same data is Pearson's Chi-Square Test for "goodness of fit."* As stated above, the essential problem is to compare the obtained distribution with a normal distribution of the same area, and of the same mean and σ , in order to determine whether the differences between the two distributions are greater than can be attributed to "chance." The quantity χ^2 (chi-square), is a generalized expression of the differences between the two distributions — theoretical and actual. Entering the table on page 124 with the value of χ^2 and n' , the number of step-intervals in the distribution, we are able to find whether the divergence of the

* Pearson, Karl, *Tables for Statisticians and Biometricians*, 1914, Introduction xxxi-xxxii, and pp. 26-28.

TABLE 17
FREQUENCY DISTRIBUTION OF 206 FRESHMEN UPON THE
THORNDIKE INTELLIGENCE EXAMINATION

Scores	<i>f</i>	
115-119	1	
110-114	2	
105-109	4	
100-104	10	Mean = 81.59
95-99	13	$\sigma = 12.14$
90-94	18	
85-89	34	
80-84	30	
75-79	37	
70-74	27	
65-69	15	
60-64	10	
55-59	2	
50-54	2	
45-49	1	
<i>N</i> = 206		

given distribution from the ideal normal form is — or is not — significant.

The first step in making the Chi-Square Test is to compute the normal frequencies as shown in Table 18. The data selected for illustration have been taken from Table 17 above, and represent the scores made on the Thorndike Intelligence Examination by 206 college freshmen. The mean test score is 81.59 and the σ is 12.14. In Table 18, column (1), the *upper limits* of the step-intervals have been listed in order. The entries in column (2) are the deviations of these upper limits from the mean. Thus, 119.5 (upper limit of step 115-119) - 81.59 (mean) = 37.91; 114.5 - 81.49 = 32.91, and so on. Each entry following is positive (and decreases by 5) down to the entry opposite 79.5. Beginning with 79.5, the deviations are negative down to 44.5, upper limit of step 40-45, whose deviation from the mean is - 37.09. In column (3), each deviation from the mean has been divided by 12.14, and is thus expressed in σ -terms; and in column (4), the percents of the area of a normal curve lying between the mean and the given σ -values have been entered. There is, for example, 49.91% of

TABLE 18
SHOWING CALCULATION OF NORMAL FREQUENCIES
(Data from Table 17, p. 120)

(1)	(2)	(3)	(4)	(5)	(6)
Upper Limit Step-Intervals	Deviations from Mean	Deviations in σ Units	Area from Mean	Area between Step- Limits (Percents)	Theoretical Frequency. Last Value $\times 206$
119.5	37.91	3.12	.5000	.0009	.19
114.5	32.91	2.71	.4991	.0025	.52
109.5	27.91	2.30	.4966	.0073	1.50
104.5	22.91	1.89	.4893	.0187	3.85
99.5	17.91	1.48	.4706	.0400	8.24
94.5	12.91	1.06	.4306	.0752	15.49
89.5	7.91	.65	.3554	.1132	23.32
84.5	2.91	.24	.2422	.1474	30.36
79.5	- 2.09	-.17	.0948	.1623	33.43
74.5	- 7.09	-.58	.0675	.1515	31.21
69.5	- 12.09	- 1.00	.2190	.1223	25.19
64.5	- 17.09	- 1.41	.3413	.0794	16.36
59.5	- 22.09	- 1.82	.4207	.0449	9.25
54.5	- 27.09	- 2.23	.4656	.0215	4.43
49.5	- 32.09	- 2.64	.4871	.0088	1.81
44.5	- 37.09	- 3.06	.4959	.0030	.62
Below 44.5			.4989	.0011	.23
			.5000		
				<u>1.0000</u>	<u>206 00</u>
	Mean = 81.59				
	$\sigma = 12.14$				

the area of a normal curve between the mean and 3.12σ (Table 15); 49.66% of the area between the mean and 2.71σ , and so on. The percents of area between the mean and the negative σ -values are read from Table 15 in the same way as are the positive σ -values.

In column (5), the percents of area between the upper and lower limits of each step-interval have been calculated. These percents are found as follows: .4991 of the area in a normal curve lies between the mean and 3.12σ , and .4966 lies between the mean and 2.71σ . Hence, $.4991 - .4966$ or .0025 of the total area lies between 3.12σ and 2.71σ or between 119.5 and 114.5. The percentage of area within the other step-intervals

is found in the same way, by subtracting the next lower percent from the percent opposite the given upper limit of a step. For example, $.4966 - .4893 = .0073$; $.4893 - .4706 = .0187$; and so on down the column. One change in procedure is necessary at the center of the distribution, where the sign of the σ -deviation changes from plus to minus. The entry .1623 (percentage of area between 79.5 and 84.5) is the sum of .0948 and .0675; that is, the percentage of the area between the mean and $.24\sigma$ is added to the percentage of area between the mean and $-.17\sigma$ to give the total percentage of area lying between 79.5 and 84.5. It will be noted that .0009 of the distribution lies *above* 119.5, upper limit of the top interval, and .0011 lies *below* 44.5, lower limit of the bottom interval. If these two small fractions of area are included in column (4) we account for the 50% of the distribution above, and the 50% of the distribution below, the mean. Hence, .5000 may be entered as the final "area-from-the-mean" percentage at the top and at the bottom of column (4); and the total area of the distribution (i.e., total N) will now add up to 100% exactly.

In the last column, column (6), each percent frequency (i.e., percentage of the distribution lying between step-limits) has been multiplied by 206, the N of the sample. This operation gives the frequencies of the best fitting normal curve of the same area, same σ and same mean, as the given distribution. In order to take account of the two small percentage frequencies which lie beyond the bounds of our distribution, we have combined the one with the frequency in the first interval, and the other with the frequency in the last interval. The first interval, therefore, now has a frequency of .71, and the last interval, a frequency of .85.

The theoretical or normal frequencies (f) calculated in Table 18 have been compared with the obtained frequencies (f_o) in Table 19. As shown in the table, each theoretical f is subtracted from its corresponding obtained f_o ; these differences ($f_o - f$) are then squared and each divided by its f . For example, the $(f_o - f)^2$ of .0841 is divided by .71, to give .118;

TABLE 19
CALCULATION OF χ^2
(Data from Tables 17 and 18)

Scores	f_o	f	$f_o - f$	$(f_o - f)^2$	$\frac{(f_o - f)^2}{f}$
115-119	1	.71	+.29	.0841	.118
110-114	2	1.50	+.50	.2500	.167
105-109	4	3.85	+.15	.0225	.006
100-104	10	8.24	+1.76	3.0976	.376
95-99	13	15.49	-2.49	6.2001	.400
90-94	18	23.32	-5.32	28.3024	1.214
85-89	34	30.36	+3.64	13.2496	.436
80-84	30	33.43	-3.43	11.7649	.352
75-79	37	31.21	+5.79	33.5241	1.074
70-74	27	25.19	+1.81	3.2761	.130
65-69	15	16.36	-1.36	1.8496	.113
60-64	10	9.25	+.75	.5625	.061
55-59	2	4.43	-2.43	5.9049	1.333
50-54	2	1.81	+.19	.0361	.020
45-49	1	.85	+ .15	.0225	.027
$N = 206$	<u>206.00</u>	<u>0.00</u>		$\chi^2 = 5.827$	
$n' = 11$ $\chi^2 = 5.827$	$P = .83$ (by interpolation in Table 20)				

.25 is divided by 1.5 to give .167, etc. The sum of the last column gives χ^2 . The formula for χ^2 (chi square), the measure of goodness of fit, may be written

$$\chi^2 = \sum \left(\frac{(f_o - f)^2}{f} \right) \quad (21)$$

(chi-square formula for testing goodness of fit)

In the present problem, $\chi^2 = 5.827$, and n' , the number of step-intervals, is 11.* Entering Table 20, we find by interpolation between 5 and 6 that $P = .83$. This P (or probability) is interpreted to mean that there are 83 chances in 100 that the given discrepancies (i.e., differences) between obtained and normal frequencies are attributable to fluctuations of random sampling. Our "fit," therefore, is exceptionally close and the given distribution may be treated as normal.

* The three small frequencies at the two extremes of the distribution have been combined, since the χ^2 test is more valid when no theoretical f is less than 5.

TABLE 20
VALUES OF P FOR TESTING GOODNESS OF FIT *

χ^2	n'								
	7	8	9	10	11	12	13	14	15
1	.986	.995	.998	.999	1.000	1.000	1.000	1.000	1.000
2	.920	.960	.981	.991	.996	.998	.999	1.000	1.000
3	.809	.885	.934	.964	.981	.991	.996	.998	.999
4	.677	.780	.857	.911	.947	.970	.983	.991	.995
5	.544	.660	.758	.834	.891	.931	.958	.975	.986
6	.423	.540	.647	.740	.815	.873	.916	.946	.966
7	.321	.429	.537	.637	.725	.799	.858	.902	.935
8	.238	.333	.433	.534	.629	.713	.785	.844	.889
9	.174	.253	.342	.437	.532	.622	.703	.773	.831
10	.125	.189	.265	.350	.440	.530	.616	.694	.762
11	.088	.139	.202	.276	.358	.443	.529	.611	.686
12	.062	.101	.151	.213	.285	.363	.446	.528	.606
13	.043	.072	.112	.163	.224	.293	.369	.448	.527
14	.030	.051	.082	.122	.173	.233	.301	.374	.450
15	.020	.036	.059	.091	.132	.182	.241	.307	.378

Ordinarily, a P must be .02 or less before the obtained distribution is considered to deviate significantly from the normal form. When $P = .02$, there are only 2 chances in 100, or 1 in 50, that the discrepancies between the frequencies of the obtained and normal distributions could have arisen by "chance," that is, by fluctuations of sampling. In such cases, therefore, the evidence is conclusive that the obtained distribution is not normal.

4. Comparing a Given Distribution with the Best Fitting Normal Curve

The frequency distribution representing the scores of 20 freshmen on the Thorndike Intelligence Examination has been plotted in Figure 24; and over it, on the same axes, is plotted the best fitting normal curve describing the same data. The obtained scores are plotted in the form of a histogram in order to bring out more clearly their apparent normality of distribution.

* This table was taken from a more complete table of P and χ^2 ; see Pearson, K., *Tables for Statisticians and Biometricians*, 1914, pp. 26-28. Note that n' in this table equals $(n + 1)$ in Table 52, p. 379.

tion when compared with the theoretical curve. As we have found from the Chi-Square Test in the last section, the correspondence between the two distributions is very close and the given data are essentially "normal" in form. Inspection of the curves "by eye" certainly confirms this conclusion.

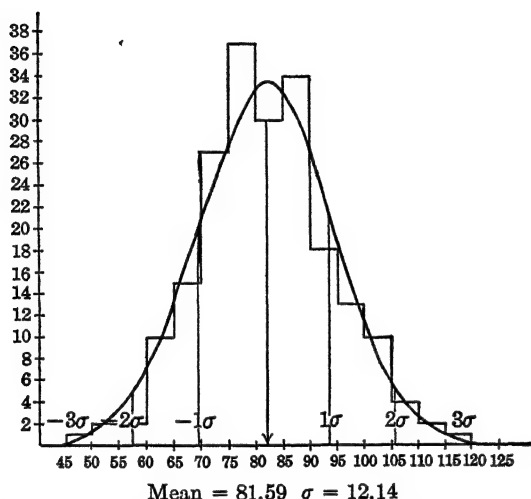


FIG. 24. Frequency Distribution of the Scores of 206 Freshmen on the Thorndike Intelligence Examination, Compared with Best Fitting Normal Curve for Same Data. (For data, see Table 17)

NORMAL CURVE ORDINATES AT MEAN, $\pm 1\sigma$, $\pm 2\sigma$, $\pm 3\sigma$

$$\begin{aligned}
 y_0 &= \frac{N}{\sigma\sqrt{2\pi}} = \frac{206}{2.51 \times 2.43} = 33.8 \\
 \pm 1\sigma &= .60653 \times 33.8 = 20.5 \\
 \pm 2\sigma &= .13534 \times 33.8 = 4.6 \\
 \pm 3\sigma &= .01111 \times 33.8 = .4
 \end{aligned}$$

The normal curve in Figure 24 has been plotted from the theoretical or normal frequencies which were determined in Table 18. It is possible, however, to plot a normal curve very simply without first determining the normal frequencies as was done in Table 18. This simple method is useful when one wishes to plot over an obtained distribution a normal distribu-

tion of the same area, mean, and σ , in order to judge goodness of fit "by eye"; or when one wishes to construct a normal curve for any purpose. The procedure is illustrated in Figure 24. First, we calculate the height of the maximum ordinate (y_0) or the frequency at the mean. This value may be determined from the equation of the normal curve given on page 108. When x in this equation is put equal to 0 (the x at the mean of the normal curve is 0), $y_0 = \frac{N}{\sigma\sqrt{2\pi}}$. In the present

problem, $N = 206$, $\sigma = 2.43$ (in *units of step-interval*), and $\sqrt{2\pi} = 2.51$. Hence, by substitution, $y_0 = 33.8$. When we know y_0 , the height of any other ordinate may be found from Table 21. This table gives the heights of the ordinates in a normal probability curve (at various σ -distances from the mean) expressed as fractions of the height of the maximum ordinate at the mean (i.e., y_0) which is taken as unity. To find, for instance, the height of the ordinate at $\pm 1\sigma$, we take the

entry .60653 from the table, opposite $\frac{x}{\sigma} = 1.0$. This entry means that when the maximum ordinate (y_0) is taken as 1.00000, the ordinate which is $\pm 1\sigma$ removed from the mean is .60653. This ordinate, in other words, is always about 60% of the height of the maximum ordinate erected at the center of the distribution. Hence, to get the height of the ordinate in the present problem which is 1σ from the mean, we simply multiply 33.8 by .60653 to get 20.5. The heights of the ordinates at $\pm 2\sigma$, and $\pm 3\sigma$ are found in the same way. (See Fig. 24 for calculation.) The entries (.13534 and .01111)

are taken from Table 21 opposite 2.0, and 3.0 in the $\frac{x}{\sigma}$ column,

and multiplied in each case by 33.8. The normal curve may be sketched in with fair accuracy from these seven points. Somewhat greater accuracy may be obtained if intermediate points, $\pm 1.5\sigma$, and $\pm 2.5\sigma$, for example, are also calculated; but seven points are usually sufficient to give the outline of the curve.

The height of the ordinate erected at the mean can be computed from $y_0 = \frac{N}{\sigma\sqrt{2\pi}}$ where $\sqrt{2\pi} = 2.51$. The height of any other ordinate, in terms of y_0 , can be read from the table when one knows the distance which the ordinate is from the mean. For example, the height of an ordinate at a distance of 1.50 σ from the mean is 32465 y_0 ; the height of an ordinate at a distance of -2.37 σ from the mean is 06029 y_0 .

IV. WHY OBTAINED DISTRIBUTIONS OFTEN DEVIATE
FROM THE NORMAL FORM

Many distributions which we encounter in practice deviate more or less markedly from the normal form. There are many reasons why this is true, some of which may be discussed here. One should hardly expect the distribution of I.Q.'s obtained from a group of 25 eight year old boys to be normal; nor the distribution of I.Q.'s obtained from a special class made up of a few dull-normal and many feeble-minded children to be normal even though the group were fairly large. The small size of the group in the first case, and the obvious special selection * in the second, are sufficient to account for the skewness or kurtosis present. Again, technical faults in the construction of a test, errors in scoring and the like, may produce skewness in an essentially symmetrical distribution of test scores.

In addition to these fairly obvious causes of asymmetry, skewness and kurtosis may appear because of a real lack of normality in the data.† True "non-normality" of distribution will arise, for instance, when one or more of the factors determining a set of scores is dominant or prepotent over the others, and hence is present more often than "chance" would allow. A simple illustration may be found in those distributions which result from the throwing of loaded dice. When off-center dice are thrown, and the results tabulated, the distribution will ordinarily be skewed and probably peaked, due to the greater potency of certain faces. Again, to take an illustration dealing with empirical data, a graph representing the frequency of automobile accidents in New York City from Monday to

* A selected group is one which is not truly representative of the larger group or population from which it is drawn. (See Chapter VIII, p. 244.)

† Theoretically, there is no real reason why distributions should always be normal. Thorndike has written: "There is nothing arbitrary or mysterious about variability which makes the so-called normal type of distribution a necessity, or any more rational than any other sort, or even more to be expected on *a priori* grounds. Nature does not abhor irregular distributions." — *An Introduction to the Theory of Mental and Social Measurements*, 1913, pp. 88-89.

Sunday will exhibit both negative skewness and leptokurtosis, because of the marked increase in traffic over the week-end.

Many illustrations of true "non-normality" of distribution could be cited also from the field of mental measurement. If an arithmetic test which requires only the four fundamental operations is given to 1000 eighth grade children, there will be a piling up of scores toward the high score end of the distribution, a negative skewness. On the other hand, if the test contains only problems in fractions, square root, interest, and the like, there will be a piling up of scores (and a shift in the peak of the curve) toward the low score end of the scale — a positive skewness. Both distributions will probably be more peaked (leptokurtic) than the normal. Such results as these may be explained in terms of the small positive and negative factors which determine the probability distribution. Too easy a test excludes from operation some of the factors which make for an extension of the curve at the upper end of the scale, such as knowledge of more advanced arithmetical relations which the brighter children would know. Too hard a test excludes from operation factors which make for the extension of the curve at the lower end of the scale, such as a knowledge of those very simple facts which would have permitted the answering of a few, at least, of the easier questions had any of these been included. In the one case we have a number of perfect scores and little discrimination; in the second case, a number of zero scores and equally poor discrimination.*

Asymmetry or distortion of distribution may be brought about by many other causes. Among the more important may be mentioned differential rates of growth; wide differences in practice or in opportunities for learning; differences in socio-economic status, deprivations, etc.†

* For a number of distributions skewed because of the causes described, see Brigham, C. C., *A Study of American Intelligence*, 1923, Section I

† Thorndike, E. L., et al., *The Measurement of Intelligence*, 1926, Chapter VIII.

PROBLEMS

1. In two throws of a coin, what is the probability of throwing a head?
2. What is the probability of throwing exactly one head in three throws of a coin?
3. Five coins are thrown. What is the probability that exactly two of them will be heads?
4. If the probability of answering a certain question correctly is four times the probability of answering it incorrectly, what is the probability of answering it correctly?
5. A rat has five choices to make of alternate routes in order to reach the food-box. If it is true that for each choice the odds are 2 to 1 in favor of the correct pathway, what is the probability that the rat will make all of his choices correctly?
6. Assume that trait X is completely determined by 6 factors — all similar and independent, and each as likely to be present as absent — plot the distribution which one might expect to get from the measurement of trait X in an unselected group of 1000 people.
7. What percentage of a normal distribution is included between the

(a) mean and 1.54σ	(d) $-3.5PE$ and $1.0PE$
(b) mean and $-2.7PE$	(e) $.66\sigma$ and 1.78σ
(c) -1.73σ and $.56\sigma$	(f) $-1.8PE$ and $-2.5PE$
8. Compute measures of skewness and kurtosis for each of the four frequency distributions in Chap. II, Problem 1, page 30.
9. (a) Test the frequency distribution given in Chap. II, Problem 1(4), page 30, for normality using the χ^2 test.
 (b) Fit a normal probability curve to the same distribution, using the method given on page 125.

ANSWERS

- | | | | | |
|------------------|------------------|--------------------|------------------|---------------------|
| 1. $\frac{3}{4}$ | 2. $\frac{3}{8}$ | 3. $\frac{10}{32}$ | 4. $\frac{4}{5}$ | 5. $\frac{32}{243}$ |
|------------------|------------------|--------------------|------------------|---------------------|
-
- | | |
|--------------|-----------|
| 7. (a) .4383 | (d) .7409 |
| (b) .4657 | (e) .2171 |
| (c) .6705 | (f) .0665 |

8.	<i>Skewness</i>		<i>Kurtosis</i>
	By formula (18)	By formula (19)	By formula (20)
(1)	— .018	— .27	.239
(2)	.156	1.03	.277
(3)	.071	.55	.222
(4)	.032	— .35	.248

9. (a) $\chi^2 = 3.14$; $n' = 11$; $P = .98$

CHAPTER VI

APPLICATIONS OF THE NORMAL PROBABILITY CURVE

I. PROBLEMS INVOLVING PERCENTAGES OF AREA WITHIN DIFFERENT PARTS OF THE NORMAL DISTRIBUTION

THIS section will consider a variety of problems which may be readily solved if we can assume that the distribution of scores with which we are dealing can be treated as normal, or at least approximately normal, in form. Each general problem will be illustrated by several examples. These examples are intended to present the issues concretely, and should be carefully worked through by the student. Constant reference will be made to Tables 15 and 16; and a knowledge of how to use these tables is essential.

1. To Determine the Percentage of Cases in a Normal Distribution Which Falls Within Given Limits

Problem (1) Given a normal distribution with a mean of 12, and a σ of 4. (a) What percentage of the cases falls between 8 and 16? (b) What percentage of the cases lies above 18? (c) Below 6?

(a) A score of 16 is four points above the mean, and a score of 8 is four points below the mean. If we divide this scale distance of four score units by the σ of the distribution (i.e., by 4) it is clear that 16 is 1σ above the mean, and that 8 is 1σ below the mean (see Fig. 25, p. 133). There are 68.26% of the cases in a normal distribution between the mean and $\pm 1\sigma$ (Table 15). Hence, 68.26% of the scores in this distribution, or approximately the middle two-thirds, fall between 8 and 16. This result may also be stated in terms of "chances." Since 68.26% of the cases in the given distribution

In a normal distribution $Q = PE$. A score of 22 is 7.75 units, or $-1.70 PE$ ($7.75/4.56 = 1.70$) from the mean; and a score of 26 is 3.75 or $-.82 PE$ from the mean (Fig. 26, below). From Table 16 we find that 37.42% of the cases in a normal distribution lie between the mean and $-1.70 PE$; and that 20.99% (by interpolation) of the cases lie between the

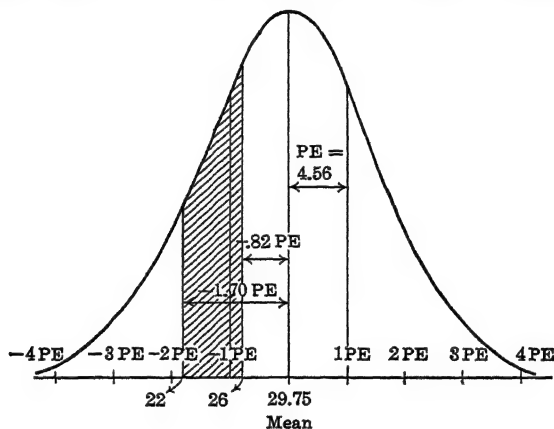


FIG. 26.

mean and $-.82 PE$. By simple subtraction, therefore, 16.43% of the cases fall between $-1.70 PE$ and $-.82 PE$ or between score 22 and score 26. The chances are about 16 in 100 that a score will fall between 22 and 26.

2. To Find the Limits in Any Normal Distribution which Will Include a Given Percentage of the Cases

Problem (1) Given a normal distribution with a mean of 16.00 and a σ of 4.00. What limits will include the middle 75% of the cases?

The middle 75% of the cases in a normal distribution must include the 37.5% just above, and the 37.5% just below the mean. From Table 15 we find that 3749 cases in 10,000, or 37.5% of the distribution, fall between the mean and 1.15 σ ; and, of course, 37.5% of the distribution also fall between the

mean and -1.15σ . The middle 75% of the cases, therefore, lie between the mean and $\pm 1.15\sigma$; or, since $\sigma = 4.00$, between the mean and ± 4.60 score units. Adding ± 4.60 to the mean (to 16.00), we find that the middle 75% of the scores in the given distribution lie between 20.60 and 11.40 (see Fig. 27, below).

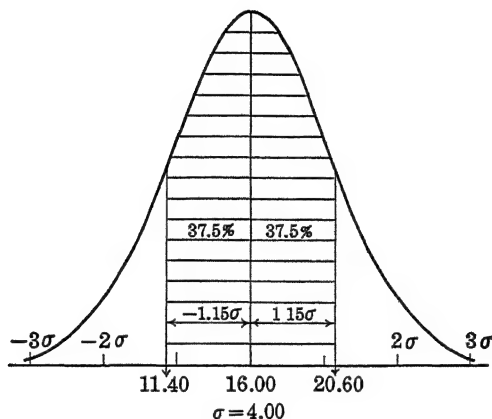


FIG. 27.

Problem (2) Given a normal distribution with a median equal to 150.00 and a Q equal to 26.00. What limits will include the *highest* 20% of the distribution? The *lowest* 10%?

The highest 20% of a normally distributed group will have 30% of the cases between its lower limit and the median, since 50% of the cases lie in the right half of the distribution. From Table 16, we find that 3004 cases in 10,000, or 30% of the distribution, fall between the median and $1.25PE$. Since the PE of the given distribution is 26.00, $1.25PE$ will be 1.25×26.00 or 32.5 points *above* the median, namely, at 182.5. The lower limit of the highest 20% of the given group, therefore, is 182.5; and the upper limit is the highest score in the distribution, whatever that may be.

The lowest 10% of a normally distributed group will have 40% of the cases between the median and its upper limit.

Exactly 40% of the distribution fall between the median and $-1.90PE$. The PE of the given distribution is 26.00; hence, $-1.90PE$ will be 1.90×26.00 or 49.4 score units *below* the median, that is at 100.6. The upper limit of the lowest 10% of scores in the given group, therefore, is 100.6; and the lower limit is the lowest score in the distribution.

3. To Compare Two Distributions in Terms of "Overlapping"

Problem (1) Given the distributions of the scores made on a logical memory test by 300 boys and 250 girls (Table 22). The boys' mean score is 21.49 with a σ of 3.63. The girls' mean score is 23.68 with a σ of 5.12. The medians are: boys, 21.41, and girls, 23.66. What percentage of boys reaches or exceeds the "median girl" (i.e., the median score in the girls' distribution)?

On the assumption that these distributions are sensibly normal, we may solve this problem by means of Table 15. The girls' median score is $23.66 - 21.49$ or 2.17 score units above the boys' mean. Dividing 2.17 by 3.63 (the σ of the boys' distribution), we find that the girls' median score is $.60\sigma$ above the mean of the boys' distribution. Table 15 shows that 23% of a normal distribution lie between the mean or median and $.60\sigma$; hence 27% of the boys ($50\% - 23\%$) must reach or exceed the score made by the "median girl."

This problem may also be solved by direct calculation from the distributions of boys' and girls' scores without any assumption as to normality of distribution. The calculations are shown in Table 22; and it will be interesting to compare the result found by direct calculation with that obtained by use of the probability tables. The problem, first, is to find the *number* of boys whose scores reach or exceed 23.66, the girls' median, and then turn this number into a percentage of the whole group. There are 217 boys who make scores up to 23.5 (upper limit of step 19.5-23.5). The step-interval 23.5-27.5 contains 68 scores; hence there are $\frac{68}{4}$ or 17 scores *per scale unit* on this step. We wish to reach 23.66 in the boys' distribu-

TABLE 22

TO ILLUSTRATE THE METHOD OF DETERMINING OVERLAPPING
BY DIRECT CALCULATION FROM THE DISTRIBUTION

<i>Boys</i>		<i>Girls</i>	
Scores	<i>f</i>	Scores	<i>f</i>
27.5-31.5	15	31.5-35.5	20
23.5-27.5	68	27.5-31.5	35
19.5-23.5	128	23.5-27.5	73
15.5-19.5	79	19.5-23.5	68
11.5-15.5	10	15.5-19.5	41
	$N = 300$	11.5-15.5	13
	$N/2 = 150$		$N = 250$
			$N/2 = 125$
$Mdn = 19.5 + \frac{1}{128} \times 4$		$Mdn = 23.5 + \frac{1}{73} \times 4$	
$= 21.41$		$= 23.66$	
$M = 21.49$		$M = 23.68$	
$\sigma = 3.63$		$\sigma = 5.12$	

What percent of the boys reach or exceed 23.66, the median of the girls? First, 217 boys make scores *below* 23.5. The step-interval 23.5-27.5 contains 68 scores; hence, there are $\frac{2}{3}$ or 17 scores *per scale unit* on this step-interval.

The girls' median, 23.66, is .16 *above* 23.5, lower limit of step-interval 23.5-27.5. If we multiply 17 (number of scores per scale unit) by .16 we obtain 2.72 which is the distance we must go into step-interval 23.5-27.5 to reach 23.66.

Now adding 217 and 2.72, we obtain 219.72 as that part of the boys' distribution which falls *below* the point 23.66 (girls' median). N is 300; hence $300 - 219.72$ gives 80.28 as that part of the boys' distribution which lies *above* 23.66. Dividing 80.28 by 300, we find that 26.76, or approximately 27%, of the boys reach or exceed the median girl.

tion. This point is .16 of a score ($23.66 - 23.50 = .16$) above 23.5, or 2.72 ($17 \times .16$) score units above 23.5. Adding 2.72 to 217, we find that 219.72 of the boys' scores fall *below* 23.66, the girls' median. Since $300 - 219.72 = 80.28$, it is clear that $80.28 \div 300$ or 26.76% (approximately 27%) of the boys make scores equal to or greater than the median score of the girls. Comparing this result with that found above on the assumption of normality, we see that the agreement is very close. Apparently the assumption of normality of distribution for the boys' scores, made above, was justified.

The agreement between the percentage of "overlapping" found by direct calculation from the distribution, and that found by use of the probability tables will usually be close,

especially if the groups are large. When the overlapping distributions are small and not very symmetrical, however, it is safer to use the method of direct calculation since no assumption as to form of distribution is then made.

4. To Determine the Relative Difficulty of Test Questions, Problems, and Other Test Items

Problem (1) Given a test question or problem solved by 10% of a large unselected group; a second problem solved by 20% of the same group; and a third problem solved by 30%. If we assume the capacity measured by the test problems to be distributed normally, what is the relative difficulty of questions 1, 2, and 3?

Our first task is to find for question 1 a position in the distribution, such that 10% of the entire group (the percent passing) lie above, and 90% (the percent failing) lie below the given point. The highest 10% in a normally distributed group has 40% of the cases between its lower limit and the mean (see Fig 28, p. 139). From Table 15 we find that 39.97% (i.e., 40%) of a normal distribution fall between the mean and 1.28σ . Hence, question 1 belongs at a point on the baseline of the curve, a distance of 1.28σ from the mean; and, accordingly, 1.28σ may be set down as the difficulty value of this question.

Question 2, passed by 20% of the group, falls at a point in the distribution 30% above the mean. From Table 15 it is found that 29.95% (i.e., 30%) of the group fall between the mean and $.84\sigma$; hence, question 2 has a difficulty value of $.84\sigma$. Question 3, which lies at a point in the distribution 20% above the mean, has a difficulty value of $.52\sigma$, since 19.85% of the distribution fall between the mean and $.52\sigma$. To summarize our results:

Question	Passed by	σ -value	σ -difference
1	10%	1.28	—
2	20%	.84	.44
3	30%	.52	.32

The σ -difference in difficulty between questions 2 and 3 is .32, which is roughly $\frac{3}{4}$ of the σ -difference in difficulty between questions 1 and 2. The percentage difference, however, is the same in the two comparisons. It is evident, therefore, that when ability is assumed to follow the normal distribution σ and not percentage differences are the real indices of differences in difficulty.

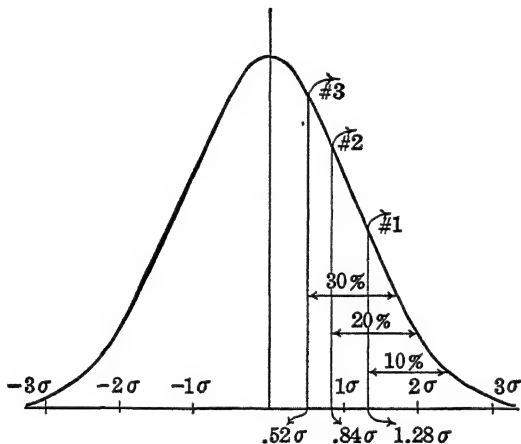


FIG. 28.

Problem (2) Given three test items, 1, 2, and 3, passed by 50%, 40%, and 30%, respectively, of a large group. Assuming normality of distribution, what percentage of this group must pass test item 4, in order for it to be as much more difficult than 3, as 2 is more difficult than 1?

An item passed by 50% of a group is, of course, failed by 50% also; and accordingly, such an item falls exactly in the middle of a normal distribution of "difficulty." Test item 1, therefore, has a σ -value of .00 since it falls exactly at the mean (Fig. 29, p. 140). Test item 2 lies at a point in the distribution 10% above the mean, since 40% of the group passed, and 60% failed this item. Accordingly, the σ -value of item 2 is .25, since from Table 15 we find that 9.87% (roughly 10%) of the cases lie be-

tween the mean and $.25\sigma$. Test item 3, passed by 30% of the group, lies at a point 20% above the mean, and this item, therefore, has a difficulty value of $.52\sigma$, as 19.85% (20%) of the normal distribution fall between the mean and $.52\sigma$.

Since item 2 is $.25\sigma$ farther along on the difficulty scale (toward the high score end of the curve) than item 1, it is clear

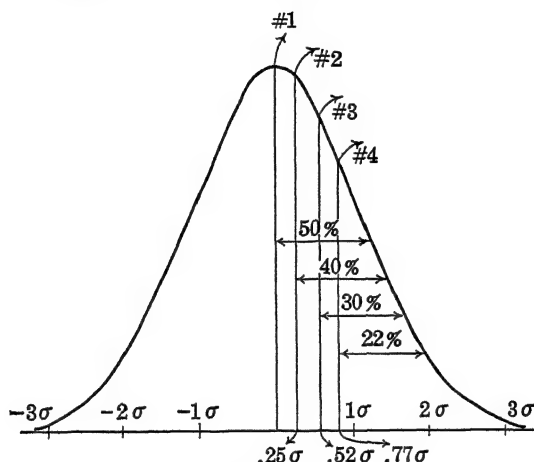


FIG. 29.

that item 4 must be $.25\sigma$ above item 3, if it is to be as much harder than item 3 as item 2 is harder than item 1. Item 4, therefore, must have a value of $.52\sigma + .25\sigma$ or $.77\sigma$; and from Table 15 we find that 27.94% (28%) of the distribution fall between the mean and this point. This means that 50% - 28% or 22% of the group must *pass* item 4. To summarize:

Test Item	Passed by	σ -value	σ -difference
1	50%	.00	—
2	40%	.25	.25
3	30%	.52	—
4	22%	.77	.25

A test item, therefore, must be passed by 22% of the group in order for it to be as much more difficult than an item passed

by 30%, as an item passed by 40% is more difficult than one passed by 50%. Note again that percentage differences are not reliable indices of differences in difficulty when the capacity measured is assumed to be distributed normally.

5. To Separate a Given Group into Sub-Groups According to Capacity, when the Capacity is Assumed to be Normally Distributed

Problem (1) Suppose that we have administered a certain examination to 100 college students. We wish to classify our group into 5 sub-groups A, B, C, D, and E according to ability, the *range* of ability to be equal in each sub-group. On the assumption that the capacity measured by our examination is normally distributed, how many students should be placed in groups A, B, C, D, and E, respectively?

Let us first represent the positions of the five sub-groups diagrammatically on a normal curve as shown in Figure 30,

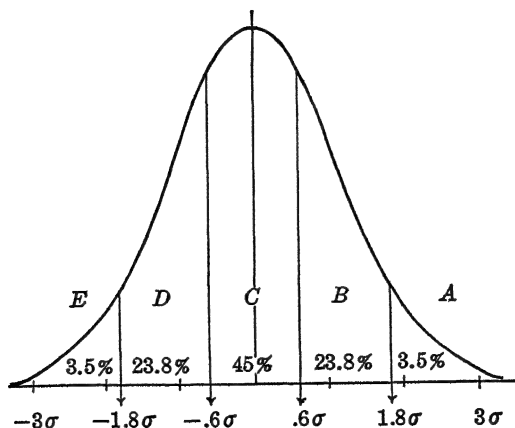


FIG. 30.

above. If the baseline of the curve is considered to extend from -3σ to $+3\sigma$, that is, over a range of 6σ , dividing this range by 5 (the number of sub-groups) gives 1.2σ as the base-

line extent to be allotted to each group. These five intervals may be laid off on the baseline as shown in the figure, and perpendiculars erected to demarcate the various sub-groups. Group A covers the upper 1.2σ ; group B the next 1.2σ ; group C lies $.6\sigma$ to the *right* and $.6\sigma$ to the *left* of the mean; groups D and E occupy the same relative positions in the lower half of the curve, as B and A occupy in the upper half.

To find what percentage of the whole group constitutes the A group, we must find what percentage of a normal distribution lies between 3σ (upper limit of the A group) and 1.8σ (lower limit of the A group). From Table 15 we read that 49.86% of a normal distribution lie between the mean and 3σ ; and that 46.41% lie between the mean and 1.8σ . Hence, 3.5% of the total area under the normal curve ($49.86\% - 46.41\%$) lie between 3σ and 1.8σ ; and, accordingly, group A comprises 3.5% of the whole group.

The percentages in the other groups are calculated in the same way. Thus, 46.41% of the normal distribution fall between the mean and 1.8σ (upper limit of group B) and 22.57% fall between the mean and $.6\sigma$ (lower limit of group B). Subtracting, we find that $46.41\% - 22.57\%$ or 23.84% of our whole group belong in sub-group B. Group C lies $.6\sigma$ above, and $.6\sigma$ below the mean. Between the mean and $.6\sigma$ are 22.57% of the normal distribution, and the same percent lies between the mean and $-.6\sigma$. Group C, therefore, includes 45.14% (22.57×2) of the whole group. Finally, sub-group D which lies between $-.6\sigma$ and -1.8σ contains exactly the same percentage of the distribution as sub-group B; and group E, which lies between -1.8σ and -3σ , contains the same percent of the whole distribution as group A. The percentages and number of men in each group are given in the following table:

	<i>Groups</i>				
	A	B	C	D	E
Percent of total in each group	3.5	23.8	45	23.8	3.5
Number in each group (100 men in all)	4 or 3	24	45	24	4 or 3

On the assumption that the capacity measured follows the normal curve, it is clear that 3 or 4 men in our group of 100 should be placed in group A, the "marked" ability group; 24 in group B, the "high average" ability group; 45 in group C, the "average" ability group; 24 in group D, the "low average" ability group; and 3 or 4 in group E, the "very low" or "inferior" group.

The above procedure may be used to determine how many students in a class should be assigned each of the five grades of A, B, C, D, or E; or it may be employed for any number of grade-groups. It must be remembered, however, that this procedure assumes that performance in the subject matter upon which the individuals are being graded is best represented by the normal curve. The larger and more unselected the group, the more nearly is this assumption justified.

II. THE SCALING OF TEST ITEMS

1. The Arrangement of Test Items into a Scale in Which the Difficulty of Each Item is Known With Reference to an Arbitrary Zero Point

- (1) Scaling Test Items in σ or PE units, when the test is intended for a single group

An important task which often confronts the worker in mental measurement is the construction of scales which shall contain problems or questions graded in difficulty from very easy to very difficult by known steps or intervals. Given a set of problems or test items, if we know what percentage of a large group passes or fails each problem it is comparatively easy to arrange the problems in a percentage order of difficulty. Such an arrangement constitutes a "scale," to be sure; but it is a very crude scale, since we know only roughly the relative difficulty of the separate items (p. 146).

In constructing scaled tests, therefore, the σ or PE of the distribution, rather than the percent passing, is taken as the unit of measurement. When the variability of the group is

employed as a scaling unit, we are able not only to arrange test items in order of difficulty but to "set" or space them at definite points along a difficulty scale. On such a scale the distance from one item to another, or from any given item to the selected zero point, is known as definitely as is the distance between two divisions on a foot-rule. To illustrate how test items are scaled when the unit of measurement is the σ or PE of the group, let us suppose that we wish to construct a scale for measuring "reasoning ability" (e.g., by means of syllogisms) in twelve year old children; or a test of arithmetic problems for Grade IV; or a scale for testing sentence memory in eight year old children. The successive steps involved in constructing such a test may be outlined as follows:

- (1) First, a large number of problems or other test items is compiled. These items should vary in difficulty from very easy to very difficult and should be representative of the field covered by the test.
- (2) Our items or problems are now given to as large and as randomly selected a group as can be assembled from among those for whom the test is eventually intended.
- (3) The percentage of the group solving each problem correctly is computed. Duplicates and items too easy or too hard, or those for one reason or another unsatisfactory, are discarded. The problems retained for the scale are then arranged in order of percentage difficulty. A problem solved correctly by 90% of the group is obviously less difficult than one solved correctly by 75%; while the second problem is, in turn, clearly less difficult than one solved correctly by 50%. The greater the percentage passing an item, the lower the position of this item in a scale of difficulty.
- (4) By means of Table 16 the percentage solving each problem correctly may now be converted into PE distances above or below the mean.* The procedure in detail is as follows. A problem solved correctly by 40% of the group is 10% or

* The procedure is identical when σ is employed instead of PE .

about $.40PE$ above the mean. A problem solved correctly by 78% of the group is 28% ($78\% - 50\%$) or $1.15PE$ below the mean. We may tabulate the results for any five items, selected at random, as follows (see Fig. 31, below):

Problems	A	B	C	D	E
Percent solving	93	78	55	40	14
Distance from mean in percentage terms	- 43	- 28	- 5	10	36
Distance from mean in PE terms	- 2.20	- 1.15	- .20	.40	1.60

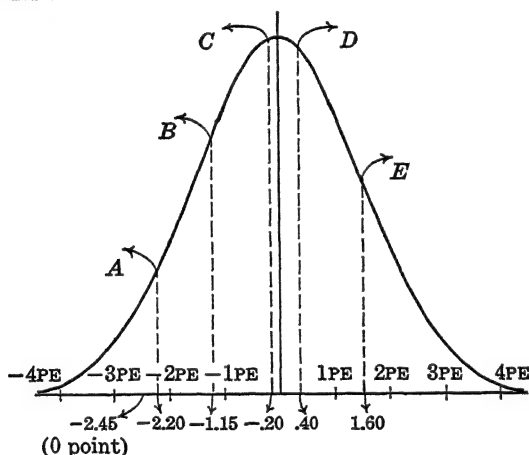


FIG. 31.

Problem A is solved by 93% of the group, i.e., by the upper 50% (the right half of the curve) plus the 43% to the left of the mean. This puts problem A at a point $- 2.20 PE$ from the mean. In the same way, the percentage distance of each problem from the mean (measured in the plus or minus direction) is first found by subtracting the percentage passing from 50%. From these percentages, the PE distance of the problem above or below the mean can be read from Table 16.*

* PE 's are taken for the percentage nearest to the given value, without interpolation.

- (5) When the *PE* distance of each problem above or below the mean has been established, the *PE* distance of each problem from the "zero point" of ability in the test may be calculated. A zero point is located in the following way. Suppose that 5% of the whole group fails to solve a single problem correctly. This would put the level of zero ability in this test 45% of the distribution below the mean, or at a point $-2.45PE$ from the mean.* The *PE* distance of each problem in the scale may now be calculated from this arbitrary zero point. To illustrate with the five problems above:

Problems	A	B	C	D	E
<i>PE</i> distance from mean . .	- 2.20	- 1.15	- .20	.40	1.60
<i>PE</i> distance from arbitrary zero, i.e., $-2.45PE$. .	.25	1.30	2.25	2.85	4.05

The simplest method of finding *PE* distances from a given zero is to subtract the zero point algebraically from the *PE* distance of each problem from the mean. Problem A, for example, is $-2.20 - (-2.45)$ or $.25PE$ from the arbitrary zero point; and Problem E is $1.60 - (-2.45)$ or $4.05PE$ from the zero point. The *PE* value of each of the other problems, as measured from the arbitrary zero point, is found in the same way. When the *PE* value-from-zero of each of the problems intended for the test has been determined, the difficulty value of each problem with respect to every other problem, as well as with respect to the arbitrary zero, is known, and the scale is finished.

A scale constructed in terms of *PE* or σ units will not ordinarily progress by equal difficulty intervals or equal units from easy items to hard items. This is inconvenient, but it does not necessarily invalidate the usefulness of the scale as a measuring instrument. In lieu of a yard-stick, one might use a measuring

* This value is an arbitrary, not a true, zero. It serves, however, as a convenient reference point (point of minimum ability) from which to measure performance.

rod upon which marks had been set at irregular intervals, say at 2.0, 3.7, 4.8 inches, with a fair degree of accuracy. But linear measurements are certainly more easily obtained with a yardstick, and scores are more easily evaluated and compared when the scale has equal units than when the units are unequal. Scale-makers, therefore, have tried as far as possible to have their scale units at least approximately equal. One method of doing this is to eliminate from the scale, as first constructed, all of the "odd" problems, retaining only those problems which fall at points approximately the same distance apart. Another method is to try out a new set of problems and from among these select those which will fill in the gaps in the scale; or change the wording or scoring of a problem in such a way as to shift it up or down on the difficulty scale.

An example of the first method of securing equal steps on a scale is furnished by Series B of the Woody Arithmetic Scales.* Woody's Arithmetic Scales cover the four fundamental operations, addition, subtraction, multiplication, and division. Series B contains problems selected from the longer Series A to give equal difficulty steps. The A scales contain problems which are progressively more difficult, but not by equal steps. In the Division Scale, for example, problem No. 1 has a difficulty value of $1.57PE$; problem No. 2 has a difficulty value of $2.08PE$; and problem No. 3 has a difficulty value of $2.18PE$ — all measured from the arbitrary zero point, $-1.67PE$.

The PE values of the 36 problems in Series A (Division), and of the 15 problems which have been selected from this list to make up Series B, are shown in Table 23. Each problem in Series A is expressed in terms of its PE distance from the arbitrary zero point, $1.67PE$ below the Third Grade median. Except in a few instances the problems in Series B appear as a graded series from easy to hard in which the intervals from problem to problem are fairly equal. Omitting the last problem, the difficulty gap from one problem to the next is, on the

* Woody, C, *Measurements of Some Achievements in Arithmetic*, Teachers College, Columbia University, Contributions to Education, No. 80, 1920, p. 39 ff.

average, .44 PE . Just as height is measured as so many feet and inches on a linear scale so the score on the given scale is simply the number of problems solved correctly: the distance the child progresses up the "mental" scale.

TABLE 23

DIFFICULTY VALUES IN TERMS OF PE OF THE PROBLEMS IN THE
WOODY ARITHMETIC SCALE FOR DIVISION

Series A and Series B are Presented for Comparison

Problem No.	Series A PE values	Series B PE values	Series B PE differences
1	1.57	1.57	
2	2.08	2.08	
3	2.18		.51
4	2.31		
5	2.40		
6	2.46		
7	2.56	2.56	.48
8	3.05	3.05	.49
9	3.16		
10	3.20		
11	3.49	3.49	.44
12	3.59		
13	3.96		
14	4.06	4.06	.57
15	4.60	4.60	.54
16	4.67		
17	4.98	4.98	.38
18	5.16		
19	5.26	5.26	.28
20	5.31		
21	5.36		
22	5.48		
23	5.56	5.56	.30
24	5.58		
25	5.78		
26	5.91		
27	6.04	6.04	.48
28	6.43	6.43	.39
29	6.76		
30	6.83	6.83	.40
31	6.87		
32	6.88		
33	7.22		
34	7.24	7.24	.41
35	8.17		
36	8.23	8.23	.99
Mean Difference =			.44 (omitting problem No. 36)

When a scale has equal steps, the increase from 10 to 12, say, is equivalent to the increase from 12 to 14, and twice the increase from 14 to 15. The child who works eight problems is as far ahead of the child who works only four, as the second child is ahead of one who cannot work a single problem. The units on the scale are equal from division to division; but it is not possible to interpret a score on a mental scale as "so many times" another score. Unlike measures of height or weight, which are taken from an absolute zero, scores on psychological and educational tests are taken from arbitrary zero points selected by the experimenter. We may correctly say that a man 72 inches tall is twice as tall as a child who measures 36 inches. But we cannot say, by analogy, that a child who scores 5 on an educational test has doubled his achievement when he is able to score 10, since the scores on our tests are not referred to an absolute zero of "just no ability."

(2) Scaling Test Items in σ or PE units when the Test is Intended for Several Groups

The method of constructing a scale outlined above may be used for any group, grade or class. When an educational scale is designed for use with more than one group, e.g., the whole elementary school, two extensions of the method are often employed. The first, which is shorter and more summary, treats children of different ages and grades as members of a single group. The second method scales separately for age or for grade and combines results into a single scale. We shall consider these two methods in order.

(A) The σ -value* of a problem is computed from the percentage of a large sample, drawn from the entire group, who pass the problem. Procedure with regard to arrangement of items in order of difficulty and location of the zero point is identical with that already described on page 146. The assumption is made that the capacity which the scale is designed to measure is distributed normally throughout the whole group;

* The method is identical when PE is used as the scaling unit.

an assumption which is not perhaps strictly true, but is not especially violent. While not as accurate as the method now to be described, this shorter plan has certain advantages in simplicity and straightforwardness.

(B) The method of scaling separately for age and grade is as follows.

(a) The σ -value of each problem is determined for each age or for each grade separately by computing the percentage passing each problem. The scale values of very easy problems, all of which are passed by the older children, are determined exclusively by the younger groups. The scale values of the more difficult problems, those out of the reach of the younger children, are determined exclusively by the older groups. Intermediate problems have scale values which are determined by several adjacent groups.

(b) The σ -distances between adjacent age or grade medians are next computed. These values are determined by finding the percentage of children in each grade or at each age who make scores higher than (i.e., "overlap") the median score of the next higher grade or age. These percents, when turned into σ -values by means of Table 15, give the σ -distances between adjoining age or grade medians (see page 136 for method of computing overlapping).

(c) Knowing the σ -distances between grade or age medians, we may convert the σ -distance of each problem from a given grade or age median into a σ -distance from the common zero point. This process involves only simple addition. The several σ -values of a problem (as determined separately from different grade or age groups) are then averaged to give a final scale value, i.e., a final distance from the common zero point.

This method of scaling over several groups is illustrated by the Woody Arithmetic Scale, Series A, the *PE* values of whose items are shown in Table 23. This test was intended for Grades III to VIII. The more difficult problems fall entirely out of the range of third grade ability, as shown by the high *PE*

values of the problems in the upper half of the Scale. When PE distances between grade medians are known, however, these difficult problems may readily be referred to the zero point ($-1.67PE$ below the third grade median).

The method of scaling test items separately for age or grade, and later combining the scale values obtained from the different groups into a single determination, assumes that the distributions of scores at successive grade or age levels exhibit equal variability, i.e., have equal σ 's. Scale values are strictly comparable (can be averaged or added) from grade to grade or age to age *only* when this condition obtains. The variability in test scores, however, usually tends to increase with age or grade level. When this is true the method of scaling described above will give many unstable and unreliable scale values. To meet this difficulty Thurstone* has devised a method of "absolute scaling" which takes account of the differences in variability at successive age and grade levels. When test items are to be scaled over a wide range of talent (say over five or six successive ages) the absolute scale will undoubtedly give the more accurate results. But when age or grade range is not wide, either of the methods described in this section will be adequate for most purposes. Unless the groups differ considerably in age or in grade, increase in variability will be slight.

2. Scaling Total Scores on a Test: the T -Scale

In the last section, the problem of scaling separate items on a test in terms of the σ or PE of the distribution was considered. We shall now describe a method of scaling total scores, or groups of items, on a test. When scaling total scores, instead of determining the difficulty value of each separate item, we determine the difficulty value, in terms of σ or PE , to be assigned to scores representing the correct response to a varying number of test items. The final score depends, therefore, upon the difficulty value of the *total* number of questions answered

* Thurstone, L. L., *A Method of Scaling Psychological and Educational Tests*, Journal of Educational Psychology, 1925, 16, pp. 433-451.

correctly, and not upon the individual difficulty values of the items.

The method of scaling total scores, known as the *T*-scale, eliminates many of the defects of both the percentile and the age scale (p. 186). This method of scaling was devised by McCall* and first used by him in constructing a series of reading tests designed for use in the grades. The original *T*-scale was drawn up from the scores made by 500 twelve year old children upon a reading test; and the scores made by other age groups were evaluated in terms of twelve year old performance. Since this first use of the method, however, *T*-scaling has been used with various other groups, so that the method no longer has reference specifically to twelve year olds.

The procedure in *T*-scaling can best be explained, perhaps, by an example. We shall outline the process in a series of steps, as it would be applied to the task of scaling a test of 20 items, administered to 200 persons.†

(1) A large and representative number of test items is compiled, varying in difficulty from very easy to very hard. These items are administered to an unselected sample of those subjects (children or adults) for whom the test is intended eventually.

(2) The percentage passing each item is computed. These percents may be converted into σ -units, so that the 20 items selected for inclusion in the final test may be chosen on the basis of their difficulty in terms of σ . Or, since an accurate measure of the difficulty of the separate items is not especially important, the 20 items in the final test may be arranged in order on the basis of percentage difficulty (number passing). The items should, of course, range from easy to hard.

(3) The test of 20 items is now administered to a representative group and the number of correct answers, i.e., total score, for each child tabulated. These total scores are then scaled as

* McCall, W. A., *How to Measure in Education*, 1929, Chap. X, pp. 272-306.

† The first two steps in *T*-scaling repeat the method of scaling separate items, outlined on page 144. These steps are included here for the sake of continuity in procedure.

shown in Table 24 for a group of 200 persons. In column (1) of the table are entered the number of items answered correctly. In column (2) are the number of subjects who answered correctly the various numbers of items listed in column (1). Three children, for example, answered no items correctly, two children answered only one item correctly, one answered two items cor-

TABLE 24

TO ILLUSTRATE THE CALCULATION OF *T*-SCORES

(1)	(2)	(3)	(4)	(5)
Total Number of Questions Answered Correctly	Number of Sub- jects Who An- swered Each Total Number of Items Cor- rectly	Number Ex- ceeding Plus One-half Reaching	Percent Ex- ceeding Plus Half Those Reaching	<i>T</i> -Score
0	3	198.5	99.25	26
1	2	196.0	98.00	29
2	1	194.5	97.25	31
3	3	192.5	96.25	32
4	5	188.5	94.25	34
5	4	184.0	92.00	36
6	6	179.0	89.50	37
7	12	170.0	85.00	40
8	20	154.0	77.00	43
9	18	135.0	67.50	45
10	25	113.5	56.75	48
11	30	86.0	43.00	52
12	10	66.0	33.00	54
13	18	52.0	26.00	56
14	10	38.0	19.00	59
15	6	30.0	15.00	60
16	12	21.0	10.50	63
17	7	11.5	5.75	66
18	5	5.5	2.75	69
19	2	2.0	1.00	73
20	1	.5	.25	78
	<u>200</u>			

rectly, three answered a total of three items correctly, and so on. In column (4) is shown the percentage exceeding plus one-half those reaching each item-total. These entries are computed in the following way [column (3)]. There are 197 children who answered at least one (i.e., more than zero) items correctly; and three children who scored exactly zero. A score of zero is an *interval* not a point (p. 2). If the three scores of 0 are dis-

tributed evenly over this interval, $1\frac{1}{2}$ scores will fall *below*, and $1\frac{1}{2}$ scores will fall *above* its midpoint. Adding $1\frac{1}{2}$ scores to 197, therefore, we obtain 198.5, the total number of scores *above* the midpoint, or the number exceeding plus one-half reaching the midpoint of score zero. When 198.5 is divided by 200, we get 99.25% as the fractional part of the whole distribution above 0, or more exactly above the *midpoint* of the zero interval.

The other entries are calculated in the same way. To take the second entry, for example, 195 (number of scores *above* 1)

TABLE 25

TO FACILITATE THE CALCULATION OF *T*-SCORES

The "percents" refer to the percent exceeding plus one-half reaching a given σ -value in the distribution. These σ -values are expressed in the form of *T*-Scores.

Percent	<i>T</i> -Score	Percent	<i>T</i> -Score	Percent	<i>T</i> -Score	Percent	<i>T</i> -Score
99.9968	10	97.72	30	50.00	50	2.28	70
99.9952	11	97.13	31	46.02	51	1.79	71
99.9928	12	96.41	32	42.07	52	1.39	72
99.9890	13	95.54	33	38.21	53	1.07	73
99.984	14	94.52	34	34.46	54	0.82	74
99.977	15	93.32	35	30.85	55	0.62	75
99.966	16	91.92	36	27.43	56	0.47	76
99.952	17	90.32	37	24.20	57	0.35	77
99.931	18	88.49	38	21.19	58	0.26	78
99.903	19	86.43	39	18.41	59	0.19	79
99.865	20	84.13	40	15.87	60	0.13	80
99.81	21	81.59	41	13.57	61	0.097	81
99.74	22	78.81	42	11.51	62	0.069	82
99.65	23	75.80	43	9.68	63	0.043	83
99.53	24	72.57	44	8.08	64	0.034	84
99.38	25	69.15	45	6.68	65	0.023	85
99.18	26	65.54	46	5.48	66	0.016	86
98.93	27	61.79	47	4.46	67	0.011	87
98.61	28	57.93	48	3.59	68	0.007	88
98.21	29	53.98	49	2.87	69	0.0048	89
						0.0032	90

plus 1 ($\frac{1}{2}$ of the scores at 1) is 196; and dividing by 200 we obtain 196/200 or 98.00 as the percentage exceeding plus one-half those reaching item 1. Again, 186 (number of scores exceeding 4) plus 2.5 ($\frac{1}{2}$ of the scores at 4) is 188.5, which gives 94.25 as the percentage exceeding plus one-half those reaching a score of 4.

(4) The percents in column (4) are readily turned into T -scores [column (5)] by means of Table 25. The T -scores in Table 25 corresponding to the percents nearest to those wanted are taken without interpolation, as fractional T -scores are a needless refinement.

The student has doubtlessly already realized that T -scores are σ -scores * multiplied by 10, and referred to an arbitrary zero point below the mean in order to avoid negative signs. We have found that in σ -scaling the mean is taken at 0 and the σ put equal to 1. The point of reference in σ -scaling, therefore, is 0 and the unit of measurement is 1. Now if the point of reference is moved from the middle of the curve (the mean) to a location -5σ below the mean, this new reference point will become 0 and the mean will become 5. The σ -divisions above the mean ($+1\sigma$, $+2\sigma$, $+3\sigma$, $+4\sigma$, and $+5\sigma$) become 6, 7, 8, 9, and 10; and the σ -divisions below the mean (-1σ , -2σ , -3σ , -4σ , and -5σ) become 4, 3, 2, 1, and 0. The σ of the distribution, of course, remains equal to 1 (see Fig. 32).

Only relatively slight changes are necessary in this σ -scale in order to make it into a T -scale. The T -scale begins at -5σ and ends at $+5\sigma$. But σ is multiplied by 10, so that the mean equals 50 and the other σ -divisions become 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100. The relationship of the T -scale to the ordinary σ -scale is shown clearly in Figure 32. Note that the T -scale ranges from 0 to 100; that its unit (T) is 1 (i.e., .1 of σ which is taken equal to 10) and that its mean is exactly 50. The zero point on the T -scale is set at -5σ in order to have the scale cover exactly 100 units. This is convenient, but it extends the extremes of the scale far beyond the ability of most groups. In actual practice, therefore, T -scores will range between 15 and 85 (i.e., from -3.5σ to $+3.5\sigma$).

While Table 25 is useful † in enabling one to calculate T -scores easily, the student should note that T -scores can also be calcu-

* See also Chapter VII, p 178.

† In Table 25 the percents lying to the *right* (above) succeeding σ -points on the baseline are tabulated, rather than percents *between* the mean and given σ -points, as in Table 15.

lated from Table 15. Let us illustrate with the score of 7, (Table 24) which has a percent exceeding plus $\frac{1}{2}$ reaching of 85. A score "passed" by 85% lies 50-85 or - 35% to the left of the mean. From Table 15 we find that we must go - 1.04 σ from the mean in order to include the 35% lying just below the mean. Since the σ of the T -scale is 10, - 1.04 σ becomes - 10.4 in terms of T ; and dropping .4 and subtracting 10

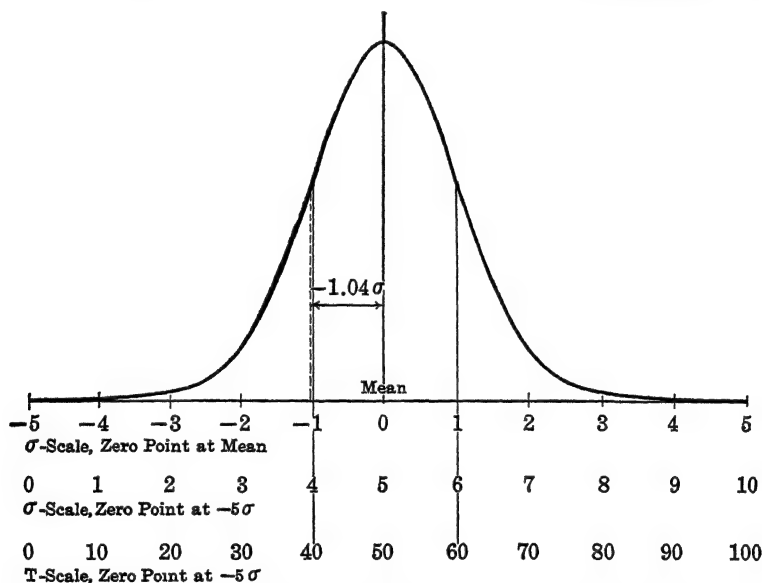


FIG. 32. To Illustrate σ -Scaling and T -Scaling in a Normal Distribution.

from 50 (the mean) we obtain 40 as the required T -score (see Fig. 32). The advantage of Table 25 over Table 15 is that one can read T -scores from the first table directly.

T -scaling is superior to the method of scaling separate items, because the difficulty value of the total score is more stable than the difficulty value of a single item. T -scores are expressed in terms of the same units and with respect to the same zero point; and unlike percentiles (p. 185) are equal throughout the scale. Hence, T -scores from different tests are directly

comparable and may be averaged or combined by simple addition. In T -scaling it is assumed that the ability measured by the test is distributed normally or approximately normally. T -scales cover a wide range of ability, and have the distinct advantage that scores from 0 to 100 are more readily understood by those possessing little knowledge of statistical method than are σ -scores.

III. THE TRANSMUTATION OF MEASURES BY RELATIVE POSITION INTO UNITS OF AMOUNT

1. Product Scales. The Conversion of Judgments of Relative Merit into σ or PE Scale Units

We have described in the last section how test items are scaled on the principle that the σ or PE value determined from the percentage passing a given item is the best index of the item's difficulty. This method is based on the assumption that ability is normally distributed on a scale from poor to good, and that performance may be scored quantitatively in terms of the amount done. It often happens, however, that the ability or trait in which we are interested is of such a nature that achievement in it cannot be measured by means of a test score representing the number of problems or questions answered correctly. This fact necessitates the construction of what are called "product scales." On such scales achievement is measured by comparing an individual's production with various "standard productions" whose values have been determined beforehand by a consensus of expert opinion. Handwriting, composition, and drawing scales are well-known examples of product scales. The excellence of a person's penmanship, for example, is evaluated by comparing a sample of his writing with various scale specimens of handwriting, the quality of which has been determined.

Product scales are constructed on the principle that "equally often noticed differences" in quality are equal. If composition A, for example, is rated better than composition B by 75% of a

group of competent judges, and composition X is rated better than composition Y by 75% of the same judges, then the difference between A and B is taken to be the same (because equally often observed) as the difference between X and Y.

Thurstone* has shown that neither the assumption that equally often noticed differences are equal, nor its converse that equal differences are equally often noticed is strictly true unless the variability of the judgments made on each specimen (the amount of agreement or disagreement) is known to be the same. Thurstone has proposed a method which takes account of the differences in the "spread" (agreement) of ratings upon different specimens; and has devised a test by means of which the equality of distribution-ratings may be determined.† The student who masters the method outlined in this section will have no difficulty in applying Thurstone's method.‡ But, before he uses the more refined technique, he should test his experimental data in order to determine whether the distributions of ratings made upon his specimens are at least approximately equal. If they are, the simpler method described below may be used without introducing error.

The assumption that "equally often noticed differences are equal" is most doubtful when applied to the scaling of items at the extremes of the qualitative range. The variability of judgments upon extremely good or extremely poor specimens will ordinarily be less than the range of judgments made upon intermediate specimens. In most product scales the accurate scaling of these extreme specimens is, perhaps, not as important as is the accurate scaling of those specimens which constitute the main body of the scale. For this reason, the assumption that equally often noticed differences are equal will usually give scale results which are as useful practically as those given by Thurstone's more refined techniques.

* Thurstone, L. L., *Equally Often Noticed Differences*, Journal of Educational Psychology, 1927, 18, pp. 289-293.

† Thurstone, L. L., *Psychophysical Analysis*, American Journal of Psychology, 1927, 38, pp. 368-389.

‡ Thurstone, L. L., *An Experimental Study of Nationality Preferences*, Journal of General Psychology, 1928, 1, pp. 405-425.

The steps in constructing a product scale, on the assumption that equally often noticed differences are equal, may be set down as follows:

- (1) A large number of specimens of the product to be scaled (e.g., handwriting samples, drawings), must first be collected. These specimens should range by gradual stages from very poor to very excellent.
- (2) A number of competent persons are asked to act as judges of the comparative excellence of the specimens. These judges are instructed to compare every specimen with every other specimen, so that a consensus of opinion on each specimen is obtained. The order of merit method, the paired comparisons method, or some variation of these, should ordinarily be employed here, as these experimental techniques provide a systematic attack upon the problem of ranking specimens for excellence
- (3) The number of times each specimen is ranked above each other specimen is now reduced to percentage terms, and these percents are expressed as σ - or PE -distances between each pair of specimens. To illustrate, if drawing A is judged better than drawing B by 65% of the group, $A-B = .39\sigma$; if B is judged better than C by 77%, $B-C = .74\sigma$. These σ -differences are read from Table 15 and are found in the following way. If a specimen is judged better than another by 50%, there is no observable difference between the two and their σ -difference is zero. But if A is judged better than B by 65%, the difference between A and B (in excess of chance) is 15%, which from Table 15 corresponds to a σ -difference of .39. In exactly the same way the difference between B and C (in excess of chance) is 27%, which corresponds to a σ -difference of .74. Figure 33 shows graphically how percentage differences can be converted into σ -differences. The distributions of judgments for A, B, and C are assumed to be normal and are taken to be equal in range and variability. The mean value of A (its scale value) is $.39\sigma$ above the mean value

of B, whose mean value is, in turn, $.74\sigma$ above the mean value of C.

- (4) When a difference has been determined for each pair of specimens, those finally selected for the scale are expressed as so many σ -units from an arbitrary zero. The procedure here may be illustrated by taking two specimens, numbers 8 and 9, from the Hillegas Composition Scale.* Hillegas

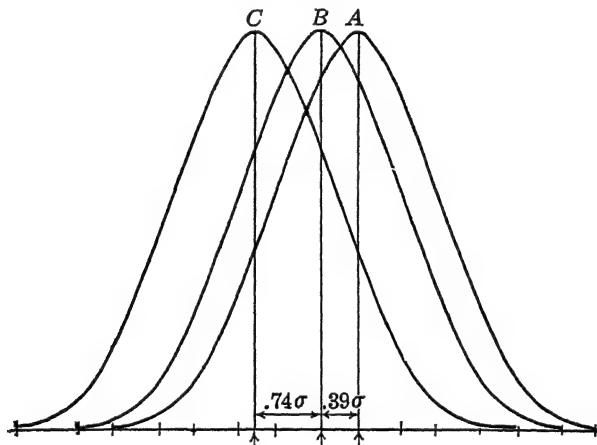


FIG. 33 To Illustrate σ -scale Differences between Specimens A, B, and C. The Distributions of Judgments on the Three Specimens are Taken to be Normal, and Equal in Range and Variability.

had each of 202 judges arrange a number of English compositions in order of merit. An artificial composition was selected as being of just zero merit, and assigned the value of 0 on the scale. Of the 202 judges, 136 or 67.33% ranked specimen 9 as better than specimen 8. From Table 16, we find that a percentage difference of 67.33 indicates a PE difference of .65, and this value expresses the amount by which 9 is better than 8. The value of specimen 8 had already been found to be $7.72PE$ above the zero point

* Hillegas, Milo B., *A Scale for the Measurement of Quality in English Composition by Young People*, Teachers College Record, 1912, 13, 4, pp. 5-55.

on the scale. Hence, specimen 9 is $7.72 + .65$ or $8.37PE$ above the zero composition. The values of the nine compositions on the Hillegas Scale as measured in PE units from the zero composition are 1.83, 2.60, 3.69, 4.74, 5.85, 6.75, 7.72, 8.37, and 9.37. Note that the steps on the scale are fairly regular and are about $1PE$ apart.

2. The Transmutation of Qualitative Data into σ -Units on the Assumption of Normality of Distribution

It is possible to represent many kinds of qualitative data in quantitative terms, if we can assume that the trait or function sampled by our data is normally distributed. Two examples, which are typical of many, will be given by way of illustration.

(1) The Scaling of Answers to a Questionnaire

The answers to the queries or statements in most questionnaires admit of several possible replies, such as Yes, No, ?; or Most, Many, Some, Few, No; or there are four or five answers one of which is to be checked. It is often desirable to "weight" these different alternatives in accordance with the degree of divergence which they indicate from the "typical answer." To do this, we must first assume that the attitude or other personality trait expressed in answering a given proposition is normally distributed. From the percentage who accept each alternative answer to a question or statement, we may then assign a σ -equivalent, which will express the value or weight to be given that answer. Lickert's* *Internationalism Scale* furnishes an example of this scaling technique. This questionnaire contains 24 statements upon each of which the subject is requested to give an opinion. Approval or disapproval of any statement is indicated by checking one of five possibilities "strongly approve," "approve," "undecided," "disapprove," and "strongly disapprove." The method of scaling as applied to statement No. 16 on the *Internationalism Scale* is shown in Table 26 on page 162. This statement reads as follows:

* Lickert, R., *A Technique for the Measurement of Attitudes*, Archives of Psychology, 1932, No. 140.

16. All men who have the opportunity should enlist in the Citizens' Military Training Camps.

Strongly approve Approve Undecided Disapprove
Strongly disapprove

The percentage selecting each of the possible answers is shown in the table.

TABLE 26

Data for Statement No. 16 of the Internationalism Scale

Answers	Strongly Approve	Approve	Undecided	Disapprove	Strongly Disapprove
Percent checking:	.13	.43	.21	.13	.10
Equivalent σ -value:	- 1.63	- .43	.43	.99	1.76

Below the percent entries are the σ -equivalents assigned to each alternative on the assumption that opinion on the question is normally distributed — that few will wholeheartedly agree or disagree, and many take intermediate views. The σ -values in Table 26 may be obtained from Table 27 (p. 164) in the following way. Reading down the first column in Table 27 headed 0, we find that beginning at the upper extreme of the normal distribution, the highest 10% of the curve has an average σ -distance from the mean of 1.76. Said differently, the average or mean of the 10% of cases at the upper extreme of the normal curve is at a distance of 1.76σ from the mean of the whole distribution. Hence, the answer "strongly disapprove" is given a weight of 1.76 (see Fig. 34).

To find the σ -value for the answer "disapprove," we select the column headed .10 and running down the column take the horizontal entry opposite 13, namely, .99. This means that when 10% of the distribution reading from the upper extreme have been accounted for, the average distance from the mean of the next 13% is $.99\sigma$. Reference to Figure 34 will make this clearer. Now from the column headed 23 (13% + 10% "used up" or accounted for), we find entry .43 opposite 21. This means that when the 23% at the upper end of the distribution

have been eliminated, the mean σ -distance from the mean of the next 21% is $.43\sigma$, which becomes the weight of the preference "undecided." The weight of the fourth answer "approve" must be found by a slightly different process. Since 44% from the upper end of the distribution have been accounted for, 6% of the 43% who marked "approve" will lie to the *right* of the mean, and 37% to the *left* of the mean, as shown in Figure 34. From the column headed 44 in Table 27, we take .08

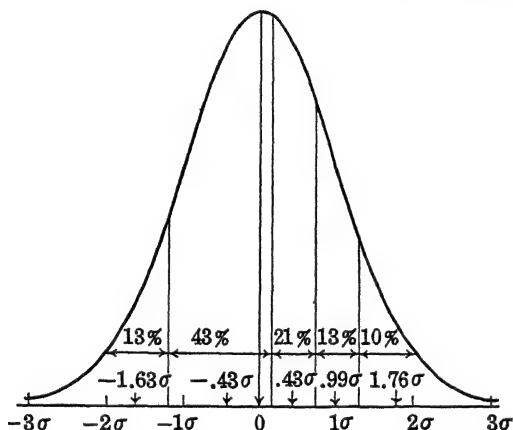


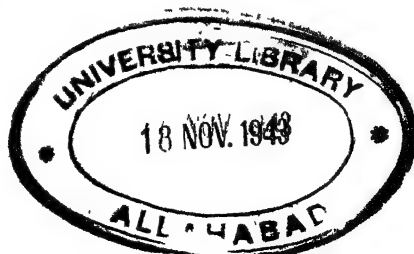
FIG. 34. To Illustrate the Scaling of the Five Possible Answers to Statement 16 on Lickert's Internationalism Scale.

(entry opposite 6%) which is the average distance from the mean of the 6% lying just above the mean. Next from the column headed 13 (50% - 37%) we take entry $-.51$, opposite 37%, as the mean distance from the mean of the 37% just below the mean. The algebraic sum
$$\frac{-.51 \times .37 + .08 \times .06}{.43} = -.43,$$

which is the weight assigned to the preference "approve." The 13% left, those marking "strongly approve," occupy the 13% at the extreme (low end) of the curve. Returning to the column headed 0, we find that the mean distance from the mean of the 13% at the extreme of the distribution is -1.63σ .

164 STATISTICS IN PSYCHOLOGY AND EDUCATION

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	270	218	196	181	170	160	151	144	137	131	125	120	115	110	106	102	97	94	90	86	82	79	76	72
2	244	207	189	175	165	156	148	141	134	128	122	118	112	108	104	99	95	92	88	84	81	77	74	71
3	228	198	182	170	160	152	144	137	131	125	120	115	110	106	102	97	94	90	86	82	79	76	72	69
4	216	191	177	165	156	148	141	134	128	123	118	113	108	104	100	96	92	88	84	81	77	74	71	67
5	210	185	172	161	152	145	138	131	126	120	115	111	106	102	98	94	90	86	82	79	76	72	69	66
6	199	179	167	157	149	141	135	129	123	118	113	108	104	100	96	92	88	84	81	77	74	71	68	64
7	192	174	163	153	145	138	132	126	121	116	111	106	102	98	94	90	86	83	79	76	72	69	66	63
8	186	170	159	150	142	135	128	124	118	113	109	104	100	96	92	88	84	81	77	74	71	68	64	61
9	181	165	155	147	139	133	126	121	116	111	106	102	98	94	90	86	83	79	76	73	69	66	63	60
10	176	161	151	143	136	130	124	119	114	109	104	100	96	92	88	85	81	78	74	71	68	65	62	59
11	171	158	148	140	134	127	122	116	111	107	102	98	94	90	87	83	79	76	73	69	66	63	60	57
12	167	154	145	138	131	125	119	114	109	105	100	96	92	89	85	81	78	74	71	68	65	62	59	56
13	163	151	142	135	128	122	117	112	107	103	99	94	91	87	83	80	76	73	70	66	63	60	57	54
14	159	147	139	132	126	120	115	110	105	101	97	93	89	85	81	78	75	71	68	65	62	59	56	53
15	156	144	136	129	123	118	113	108	103	99	95	91	87	83	80	76	73	70	66	63	60	57	54	51
16	152	141	134	127	121	116	111	106	101	97	93	89	85	82	78	75	71	68	65	62	59	56	53	50
17	149	139	131	125	119	113	109	104	99	95	91	87	84	80	77	73	70	67	64	60	57	54	52	49
18	146	136	129	122	117	111	106	102	98	93	89	86	82	78	75	72	68	65	62	59	56	53	50	47
19	143	133	126	120	114	109	105	100	96	92	88	84	80	77	73	70	67	64	61	58	55	52	49	46
20	140	131	124	118	112	107	103	98	94	90	86	82	79	75	72	69	65	62	59	56	53	50	47	45
21	137	128	121	116	110	105	101	96	92	88	84	81	77	74	70	67	64	60	58	55	52	49	46	43
22	135	126	119	113	108	103	99	95	90	87	83	79	76	72	69	66	62	59	56	53	50	48	45	42
23	132	124	117	111	106	101	97	92	89	85	81	78	74	71	67	64	61	58	55	52	49	46	43	41
24	130	121	115	109	104	100	95	91	87	83	80	76	73	69	66	63	60	57	54	51	48	45	42	39
25	127	119	113	107	102	98	93	89	85	82	78	74	71	68	64	61	58	55	52	49	46	43	41	38
26	125	117	111	105	101	96	92	88	84	80	76	73	70	66	63	60	57	54	51	48	45	42	39	37
27	123	115	109	104	99	94	90	86	82	78	75	71	68	65	62	58	55	52	49	46	44	41	38	35
28	120	113	107	102	97	92	88	84	80	77	73	70	67	63	60	57	54	51	48	45	42	39	37	
29	118	111	105	100	95	91	87	83	79	75	72	68	65	62	59	56	53	50	47	44	41	38		
30	116	109	103	98	93	89	85	81	77	74	70	67	64	60	57	54	51	48	45	42	40			
31	114	107	101	96	92	87	83	79	76	72	69	65	62	59	56	53	50	47	44	41				
32	112	105	99	94	90	86	82	78	74	71	67	64	61	58	54	51	48	46	43					
33	110	103	98	93	88	84	80	76	73	69	66	63	59	56	53	50	47	44						
34	108	101	96	91	86	82	79	75	71	68	64	61	58	55	52	49	46							
35	106	99	94	89	85	81	77	73	70	66	63	60	56	53	50	47								
36	104	97	92	88	83	80	75	72	68	65	61	58	55	52	49									
37	102	96	91	86	82	78	74	70	67	63	60	57	54	51										
38	100	94	89	84	80	76	72	69	65	62	59	55	52											
39	98	92	87	83	79	75	71	67	64	61	57	54												
40	97	91	86	81	77	73	69	66	62	59	56													
41	95	89	84	80	75	72	68	64	61	58														
42	93	87	82	78	74	70	66	63	60															
43	91	85	81	76	72	69	65	62																
44	90	84	79	75	71	67	64																	
45	88	82	78	73	69	66																		
46	86	81	76	72	68																			
47	85	79	75	70																				
48	83	78	73																					
49	81	76																						
50	80																							



APPLICATIONS OF NORMAL PROBABILITY CURVE 165

	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
1	69	66	63	60	57	54	51	48	45	43	40	37	35	32	29	27	24	21	19	18	14	11	09	06	04	01
2	67	64	61	58	55	52	50	47	44	41	39	36	33	31	28	25	23	20	18	15	13	10	08	05	03	
3	66	63	60	57	54	51	48	45	43	40	37	35	32	29	27	24	21	19	16	14	11	09	06	05		
4	64	61	58	55	52	50	47	44	41	39	36	33	31	28	25	23	20	18	15	13	10	08	05			
5	63	60	57	54	51	48	45	43	40	37	35	32	29	27	24	21	19	16	14	11	09	06				
6	61	58	55	53	50	47	44	41	39	36	33	31	28	25	23	20	18	15	13	10	08					
7	60	57	54	51	48	45	43	40	37	35	32	29	27	24	21	19	16	14	11	09						
8	58	55	52	50	47	44	41	39	36	33	31	28	25	23	20	18	15	13	10							
9	57	54	51	48	46	43	40	37	35	32	29	27	24	21	19	16	14	11								
10	56	53	50	47	44	41	39	36	33	31	28	25	23	20	18	15	13									
11	54	51	48	46	43	40	37	35	32	29	27	24	22	19	16	14										
12	53	50	47	44	41	39	36	33	31	28	25	23	20	18	15											
13	51	48	46	43	40	37	35	32	29	27	24	22	19	16												
14	50	47	44	42	39	36	33	31	28	25	23	20	18													
15	49	46	43	40	37	35	32	29	27	24	22	19														
16	47	44	42	39	36	33	31	28	26	23	20															
17	46	43	40	37	35	32	29	27	24	22																
18	44	42	39	36	33	31	28	26	23																	
19	43	40	38	35	32	30	27	24																		
20	42	39	36	34	31	28	26																			
21	40	38	35	32	30	27																				
22	39	36	34	31	28																					
23	38	35	32	30																						
24	36	34	31																							
25	35	32																								
26	34																									

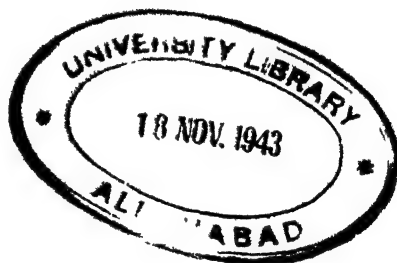


TABLE 27

Average distance from the mean, in terms of σ , of each single percentage of a normal distribution. Figures along the top of the table represent percentages of area from either extreme. Figures down the side of the table represent percentages measured from given points in the distribution.

Examples The average distance from the mean of the highest 10% of a normally distributed group is 1.76σ (entry opposite 10 in first column). The average distance from the mean of the next 20% is $.86\sigma$ (entry opposite 20 in column headed 10). The average distance from the mean of the next 30% is

$$\frac{.26 \times .20 + (-.13 \times 10)}{.30}$$

or $.13\sigma$ (20% lie to right of mean and 10% to left, see p. 163).

In order to avoid negative values each σ -weight in Table 26 can be expressed as a σ -distance from -3.00σ (or -5.00σ). If referred to -3.00σ , the weights become in order 1.37, 2.57, 3.43, 3.99, and 4.76. Dropping decimals, and taking the first two digits, we could also assign weights of 14, 26, 34, 40, and 48.

When all of the 24 statements on the Internationalism Scale have been scaled as shown above, a person's "score" (his attitude toward internationalism in general) is found by adding up the weights assigned to the various preferences which he has selected. An individual whose opinions are extreme, e.g., who tends strongly to disapprove many statements, will receive a proportionally larger total score when the choices are σ -scaled, than he would receive if the five possibilities were assigned arbitrary weights of 1, 2, 3, 4, and 5. Lickert has shown, however, that σ -scaling yields results which, *for the test as a whole*, are little if any more reliable or more discriminatory than the results obtained when the five answers are scored simply 1, 2, 3, 4, and 5. This virtual equality of theoretically more accurate and theoretically less accurate scoring methods is a rather familiar finding in mental measurement. In the present instance, it probably arises from the fact that the greater differentiation which the σ -scaling technique undoubtedly provides for *single* items is lost in the process of adding or averaging the separate scores from many items. A real advantage of σ -scaling is that the units of the scale are equal and may be compared from item to item or from scale to scale. Also, σ -scaling gives a more accurate picture of the extent to which extreme or biased opinions on a given question are really divergent from the typical opinion than does the arbitrary weighting method.

(2) The Scaling of Judgments or Ratings

In many psychological problems, individuals are rated or ranked for their possession of various traits or attributes which are not readily measured by tests. Honesty, interest in one's

work, tactfulness, originality, are illustrations of such traits. Suppose that two teachers A and B have rated a group of pupils for "social responsibility" on a 5-point scale. A rating of 1 means that the trait is possessed in marked degree, a rating of 5 that it is almost if not entirely absent, and ratings of 2, 3, and 4 indicate intermediate degrees. Assume that the percentage of children assigned each rating is as follows:

Social Responsibility		
Rating	A	B
1	10%	20%
2	15%	40%
3	50%	20%
4	20%	10%
5	5%	10%

It is obvious that B rates more leniently than A, so that a rating of 1 by B does not represent the same degree of "social responsibility" as a rating of 1 by A. Can we assign "weights" to these ratings of A and B, in accordance with the percentage of times each rating is used, so as to make the ratings of the two teachers comparable? The answer is "yes," provided we can assume that the distribution of the trait "social responsibility" is normal, and that one teacher is as competent a judge as the other. From Table 27, we may read σ -equivalents to the percents given each rating by A and B as follows:

Rating	A	B
1	1.76	1.40
2	.95	.27
3	.00	- .53
4	- 1.07	- 1.04
5	- 2.10	- 1.76

These σ -values are read from Table 27 in exactly the same way as were the σ -equivalents in the previous problem. If we assume -3.00σ as an arbitrary reference point, the σ -values for the ratings of A and B all become positive:

Rating	A	B
1	4.76	4.40
2	3.95	3.27
3	3.00	2.47
4	1.93	1.96
5	.90	1.24

Dropping decimals, and taking only the first two digits, A's and B's ratings become:

Rating	A	B
1	48	44
2	40	33
3	30	25
4	19	20
5	9	12

It is now possible to combine the ratings of A and B by adding or by averaging them. If a child receives a rating of "4" by A and a rating of "2" by B, his combined or average rating would be $\frac{-1.07 + 27}{2}$ or $-.40$; taking the σ -values referred to an arbitrary zero, $\frac{1.93 + 3.27}{2}$ or 2.60; or finally $\frac{19 + 33}{2}$ or 26.

Table 27 will prove extremely valuable in enabling us to transmute many kinds of qualitative data into quantitative terms. Almost any attribute upon which relative judgments can be obtained, may be assigned scores in a normal distribution in terms of the σ of the judgments.

(3) The Transmutation of Orders of Merit into Units of Amount

It is often desirable to transmute measures arranged in order of merit into units of amount or "scores" upon a linear scale. This may be accomplished by means of tables, provided we are justified in assuming normality for the trait in which the ranking has been made. To illustrate, suppose that 15 salesmen

have been ranked in order of merit for selling efficiency, the most efficient salesman being ranked 1, the least efficient being ranked 15. If we are justified in assuming that "selling efficiency" follows the normal probability curve, we can, with the aid of Table 28 (p. 171), assign to each man a "selling score" on a scale of 100 points. Such a score will represent his relative capacity as a salesman better than will a rank of 2, 6, or 14. The problem may be stated specifically as follows:

Problem (1) Given 15 salesmen, ranked in order of merit by their sales manager, to transmute these rankings into scores on a scale of 100 points.

First, by means of the formula

$$\text{Percent position} = \frac{100(R - .5)}{N} \quad (22)$$

(formula for transmuting ranks into percents)

in which R is the rank of the individual in the series and N is the number of individuals ranked, we determine the "percent position" of each man. From these percent positions the man's score on a scale of 100 points is then read from Table 28. Salesman A, who ranks No. 1, has a percent position of $\frac{100(1 - .5)}{15}$

or 3.33, and his score from Table 28 is 85 (finer interpolation unnecessary). In like manner, Salesman B, who ranks No. 2,

has a percent position of $\frac{100(2 - .5)}{15}$ or 10, and his score, ac-

cordingly, is 75. The scores of the other salesmen, found in exactly the same way, are given in the table on page 170.

It has been frequently pointed out that the assumption of normality in a trait implies that differences at the extremes are relatively much greater than the same differences around the mean. This is clearly brought out in the next table; for, while all differences in the order of merit series equal 1, the differences between the transmuted scores vary considerably. The greatest differences are found at the ends of the series, the smallest in the middle. For example, the difference in score between A

Salesmen	Rank	Percent Position	Score (Scale 100)
A	1	3.33	85
B	2	10.00	75
C	3	16.67	69
D	4	23.33	64
E	5	30.00	60
F	6	36.67	57
G	7	43.33	53
H	8	50.00	50
I	9	56.67	47
J	10	63.33	43
K	11	70.00	40
L	12	76.67	36
M	13	83.33	31
N	14	90.00	25
O	15	96.67	15

and B or between N and O is three times the difference between G and H. Clearly it is three times as easy for a salesman to improve sufficiently to move from eighth to seventh place, as it is for him to improve sufficiently to move from second to first place

Another use to which Table 28 may be put is in the combination of incomplete order of merit rankings. To illustrate:

Problem (2) Six persons, A, B, C, D, E, and F, are to be ranked for honesty by three judges. Judge 1 knows all six well enough to rank them, Judge 2 knows only three well enough to rank them; and Judge 3 knows four well enough to rank them. Can we obtain a fair composite order of merit ranking for all six persons by combining these three sets of rankings, two of which are incomplete?

We may tabulate our data as follows:

Persons	A	B	C	D	E	F
Judge 1's ranking	1	2	3	4	5	6
Judge 2's ranking		2		1		3
Judge 3's ranking	2		1		3	4

It seems fair that A should get more credit for ranking first in a list of six, than D for ranking first in a list of three, or C for ranking first in a list of four. In the order of merit ratings, all three individuals are given the same rank. But when we assign

TABLE 28

THE TRANSMUTATION OF ORDERS OF MERIT INTO
UNITS OF AMOUNT OR "SCORES" *

Example: If $N = 25$, and $R = 3$, Percent Position is $\frac{100(3 - .5)}{25}$ or 10
(formula 22) and from the table, the equivalent rank is 75, on a scale of
100 points.

Percent	Score	Percent	Score	Percent	Score
.09	99	22.32	65	83.31	3.1
.20	98	23.88	64	84.56	3.0
.32	97	25.48	63	85.75	2.9
.45	96	27.15	62	86.89	2.8
.61	95	28.86	61	87.96	2.7
.78	94	30.61	60	88.97	2.6
.97	93	32.42	59	89.94	2.5
1.18	92	34.25	58	90.83	2.4
1.42	91	36.15	57	91.67	2.3
1.68	90	38.06	56	92.45	2.2
1.96	89	40.01	55	93.19	2.1
2.28	88	41.97	54	93.86	2.0
2.63	87	43.97	53	94.49	1.9
3.01	86	45.97	52	95.08	1.8
3.43	85	47.98	51	95.62	1.7
3.89	84	50.00	50	96.11	1.6
4.38	83	52.02	49	96.57	1.5
4.92	82	54.03	48	96.99	1.4
5.51	81	56.03	47	97.37	1.3
6.14	80	58.03	46	97.72	1.2
6.81	79	59.99	45	98.04	1.1
7.55	78	61.94	44	98.32	1.0
8.33	77	63.85	43	98.58	.9
9.17	76	65.75	42	98.82	.8
10.06	75	67.48	41	99.03	.7
11.03	74	69.39	40	99.22	.6
12.04	73	71.14	39	99.39	.5
13.11	72	72.85	38	99.55	.4
14.25	71	74.52	37	99.68	.3
15.44	70	76.12	36	99.80	.2
16.69	69	77.68	35	99.91	.1
18.01	68	79.17	34	100.00	0
19.39	67	80.61	33		
20.93	66	81.99	32		

* From Hull, C. L., *The Computation of Pearson's r from Ranked Data*,
Journal of Applied Psychology, 1922, 6, pp. 385-390

See also Hull, C. L., *Aptitude Testing*, 1928, pp. 491-492, for a table
which converts the ranks of any series from 11 to 50 directly into units of
amount on a 10-point scale.

scores to each person, in accordance with his position in the list, by means of formula 22 and Table 28, A gets 77 for his first place, D gets 69 for his, and C gets 73 for his. See table below:

Persons	A	B	C	D	E	F
Judge 1's ranking .	1	2	3	4	5	6
score .	77	63	54	46	37	23
Judge 2's ranking		2		1		3
score . . .		50		69		31
Judge 3's ranking .	2		1		3	4
score	<u>56</u>	<u>—</u>	<u>73</u>	<u>—</u>	<u>44</u>	<u>27</u>
Sum of scores .	133	113	127	115	81	81
Average score . .	67	57	64	58	41	27
Order of Merit . . .	1	4	2	3	5	6

All of the ratings have been transmuted as shown in problem (1) above. Separate scores may be combined and averaged to give the final order of merit shown in the table.

By means of formula 22 and Table 28 it is possible to transmute any set of ranks into scores, if we may make the assumption that the trait for which the ranking is made is normally distributed. The method is useful in the case of those attributes which are not easily measured by ordinary methods, but for which individuals may be arranged in order of merit, as, for example, athletic ability, personality, beauty, and the like. It is also valuable in correlation problems when the only available criterion * of a given ability or aptitude is a set of ranks, while the tests are scored in performance units. Transmuted scores may be combined or averaged like other test scores.

A word of explanation may be added with regard to Table 28. This table represents a normal frequency distribution which has been cut off at $\pm 2.5\sigma$. The baseline of the curve is 5σ , therefore, and may conveniently be divided into 100 parts, each .05 long. The first .05 from the upper limit of the curve takes in .09 of 1% of the distribution and is scored 99 on a scale of 100. The next .05 (.10 from the upper end of the curve)

* For definition of a criterion, see Chapter XI, p. 324.

takes in .20 of 1% of the entire distribution and is scored 98. In each case, the percent position gives the fractional part of the normal distribution which lies to the right of the given σ position on the baseline. The σ -values determine the transmuted scores.

PROBLEMS

1. In a sample of 1000 cases, a certain test returns a mean of 14.40 and a σ of 2.50. Assuming normality of distribution
 - (a) How many individuals score between 12 and 16?
 - (b) How many score above 18? below 8?
 - (c) What are the chances that any individual selected at random will score above 15?
2. In a distribution of 100 cases, the median is 29.74 and the Q is 3.18. Assuming normality
 - (a) What percent of the cases lie between 24 and 25?
 - (b) What limits include the middle 60%?
 - (c) What limits include the lowest 5%?
3. In a certain achievement test, the Seventh Grade median is 28.00, with a Q of 4.80; and the Eighth Grade median is 31.60 with a Q of 4.00. What percent of the Seventh Grade is above the median of the Eighth Grade? What percent of the Eighth Grade is below the median of the Seventh Grade?
4. Two years ago a group of twelve year olds had a reading ability expressed by a mean score of 40.00 and a σ of 3.60; and a composition ability expressed by a mean of 62.00 and a σ of 9.60. To-day, the group has gained 12 points in reading, and 10.8 points in composition. How many times greater is the gain in reading than the gain in composition?
5. In Problem 2, Chap IV, we computed directly from the distribution the percent of Group A which reaches or exceeds the median of Group B. Compare this value with the percentage of overlapping obtained on the assumption of normality in Group A.
6. Four problems, A, B, C, and D, have been solved by 50%, 60%, 70%, and 80%, respectively, of a large group. Compare the difference in difficulty between A and B with the difference in difficulty between C and D.

7. In a certain college, 10 grades, A +, A, A -; B +, B, B -; C +, C, C -, and D, are assigned. On the assumption that ability in mathematics is distributed normally, how many students in a group of 500 freshmen should receive each grade?
8. Five problems are passed by 15%, 34%, 50%, 62%, and 80%, respectively, of a large unselected group. If the zero point of ability in this test is taken to be at -3σ , what is the σ -value of each problem as measured from this point?
9. In a large group of competent judges, 88% rank composition A as better than composition B; 65% rank B as better than C. If C is known to have a *PE* value of 3.5 as measured from the "zero composition," i.e., the composition of just zero merit, what are the *PE* values of B and A as measured from this zero point?
10. Twenty-five men on a football squad are ranked in order of merit from 1 to 25 for all-around playing ability by the coach. On the assumption that general playing ability is normally distributed, transmute these ranks into units of amount on a scale of 100 points
11. On an Occupational Interest Blank, each occupation is followed by five symbols, L! L ? D D!, which denote different degrees of "liking" and "disliking." The answers to one item are distributed as follows:

L!	L	?	D	D!
8%	20%	38%	24%	10%

- (a) By means of Table 27 convert these percents into σ -units.
- (b) Express each σ -value as a distance from "zero," taken at -3σ , and multiply by 10 throughout.

APPLICATIONS OF NORMAL PROBABILITY CURVE 175

12. Calculate T -scores corresponding to "number of questions answered correctly" in the following problem:

No. of Questions Answered Correctly	Students Answering Each Number	Percent Exceeding Plus One-half Reaching	T -score
0	4	99	27
1	12	95	34
2	24		
3	36		
4	40		
5	28		
6	24		
7	20		
8	6		
9	4		
10	2		
	<u>200</u>		

(The first two T -scores have been entered.)

ANSWERS

- (a) 570
(b) 75; 5
(c) 41 in 100
- (a) 5%
(b) 33.72 and 25.76
(c) 21.95 and lowest score in the distribution
- 31%; 27%
- Three times as great.
- 39% as compared with 42%.
- Difference between A and B is 25σ ; between C and D, $.32\sigma$.
- Grades: A + A A - B + B B - C + C C - D
Students
receiving 3 14 40 80 113 113 80 40 14 3
- In order: 4.04; 3.41; 3.00; 2.69; 2.16.
- B, 4.07PE; A, 5.82PE (Interpolate for B's value)
- Rank: 1 2 3 4 5 6 7 8 9 10 11 12 13
Score: 89 80 75 71 68 65 63 60 58 56 54 52 50
Rank: 14 15 16 17 18 19 20 21 22 23 24 25
Score: 48 46 44 42 40 37 35 32 29 25 20 11

176 STATISTICS IN PSYCHOLOGY AND EDUCATION

11.	L!	L	?	D	D!
(a)	- 1.86	- .94	- .08	.80	1.76
(b)	11	21	29	38	48
12.	<i>T</i> -Scores				
	27				
	34				
	39				
	44				
	49				
	54				
	58				
	62				
	67				
	71				
	76				

CHAPTER VII

COMPARABLE MEASURES; COMBINING TEST SCORES AND DISTRIBUTIONS

WHEN a number of different tests have been administered to the same subject, one often wishes (1) to compare directly the subject's standing in the various measures; or (2) to average or combine the separate test scores into a composite which will represent achievement in the test battery as a whole. In attempting to compare or to combine scores from many tests, however, difficulty arises. Mental measurements are made upon different scales and are expressed in a variety of units. Scores upon mental tests, therefore, differ markedly in the *kind* of units and the *size* of the units in which they are expressed. Time and amount scores are instances of different *kinds* of units. Tests given by the *amount-limit method* (score is the time taken to complete the task) cannot be compared directly with tests given by the *time-limit method* (score is the amount completed in a given time). Again, a height of 65 inches cannot be directly compared with a height of 147 cms.; nor can a score of 18 upon a hard reading test be directly compared with a score of 42 upon an easy reading test, owing to differences in the size of units.

Test scores expressed in the same kind of unit (e.g., time, or amount done), even if not directly comparable, may be combined into a final score. The total score obtained by adding the sub-test scores of a general intelligence test, for instance, or of an educational achievement examination, are examples of such combinations. The simplest procedure in combining scores of this sort is to add or average them. Simply to average "raw" or obtained scores, however, gives us no control over the relative importance or "weight" of the various tests in the composite

result. It is often tacitly assumed that by simply averaging test scores we avoid the troublesome question of weighting; but what we actually do, in such cases, is to weight quite drastically without knowing what the weights are. Tests which are not weighted weight themselves.

There are several methods which may be employed for reducing scores to a common basis so that they may be (1) compared directly, or (2) combined with known weights. Some of the more useful of these methods will be outlined in the present chapter.

I. METHODS OF RENDERING TEST SCORES COMPARABLE

1. Converting the Scores of Different Tests into Equivalent Units

(1) Standard or z -Scores

We have seen that the deviation of an individual's score (X) from the mean of the test (M) is represented by x ; that is, $x = X - M$ (p. 39). When the deviation (x) of an individual from the mean of the test is divided by the σ of the test, the resulting measure is variously known as a sigma-score, a "reduced" score, a standard score, or a z -score. If the scores of a subject in two tests, say, are written as standard scores, i.e., as z_1 and z_2 (these equal $\frac{x_1}{\sigma_1}$ and $\frac{x_2}{\sigma_2}$, respectively), such scores are comparable, since they are expressed in equivalent (viz., σ) units. The formula for a standard score or z -score is

$$z = \frac{X - M}{\sigma} = \frac{x}{\sigma} \quad (23)$$

(standard or z -score in a test, i.e., score expressed in σ -units)

The method of calculating and combining standard scores is illustrated for three tests in Table 29. Note that the raw scores made by Subject A on the three tests differ markedly in size. These obtained scores are not comparable as they stand; and if combined, the general intelligence test will be more

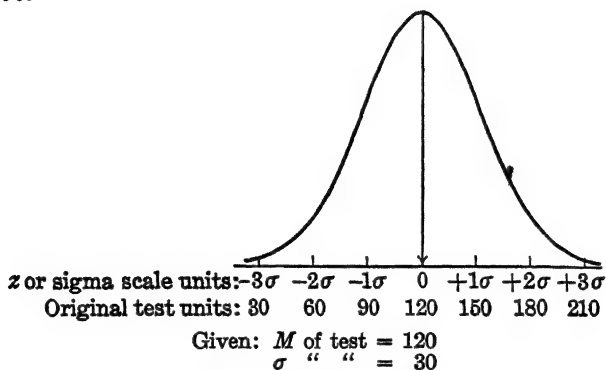
TABLE 29
TO ILLUSTRATE THE CALCULATION OF STANDARD
OR Z-SCORES FOR THREE TESTS

	Tests		
	Word-Building	Digit-Span	General Intelligence
Mean	16.3	7.4	150.4
σ	4.9	1.3	20.6
A's Scores	14	9	165
A's deviations from the mean	- 2.3	+ 1.6	+ 14.6
A's standard scores (x/σ)	- .47	1.23	.71
A's mean z-score = .49			

heavily weighted in the composite than either the digit-span or the word-building tests. Subject A's score of 14 in the word-building test is 2.3 score units *below* 16.3, the M of the test distribution. If we divide this deviation by 4.9 (the σ of the test), A's standard score is $\frac{-2.3}{4.9}$ or - .47. Subject A, therefore, stands - .47 σ below the mean of the group. On the digit-span test, A's score of 9 is 1.6 score units *above* the M , and his standard score is $\frac{1.6}{1.3}$ or 1.23. Subject A's standard score in the digit-span test is higher than his standard score in the general intelligence test; but both are above the means of their distributions. The sum of A's scores (added algebraically) is 1.47; and his average score of $\frac{1.47}{3}$ or .49 indicates that, with respect to all three tests, he ranks about $\frac{1}{2}\sigma$ above the mean of his group.

Standard scores, since they are all in σ -units, may be added, subtracted, or averaged as though they were all in inches or pounds. The student will note that what one does in calculating standard scores is to transmute the raw scores of the given test into equivalent σ -values in a "standard" normal distribution, whose mean is .00 and whose $\sigma = 1.00$. The truth of this is shown graphically in Figure 35. Note in

Figure 35 that the M of the test (i.e., 120) becomes .00 on the σ -scale; and that a score of 150, say, is equivalent to a z -score of 1.00.



Score of 150 is $+1\sigma$ from M ; hence $z = 1.0$
 " " 108 " $-.4\sigma$ " " ; " $z = -.40$

FIG. 35. How Test Scores are Transmuted into Sigma or Standard (z) Scores.

In general, two assumptions are made when z -scores are calculated and combined. The first is that the distributions of scores from different tests are at least approximately normal. The second is that a score of 1.50σ , say, in one test signifies the same degree of superiority as a score of 1.50σ in a second test. The first of these assumptions is reasonably well satisfied by most test distributions. The second, also, seems valid. Granted a normal distribution, equivalent z -scores represent at least the same degree of relative achievement with respect to the mean of the test. Since the zero point of ability in our tests is unknown, the central tendency of the group (M or Mdn) is the most logical reference point from which to measure superior or inferior performance.

(2) Equivalent Scores in a Common Distribution

There is a practical disadvantage to the use of z -scores when much calculation is to be done. Standard scores are expressed in $+$ and $-$ values, are nearly always decimals, and are often

small and inconvenient numbers to handle. For this reason it is often advisable to convert the scores of different tests into a common normal distribution in which the M and the σ are assigned specified values. Hull* has proposed a "standard" distribution with a mean of 50 and a σ of 14. Since the range in a normal distribution rarely extends more than 3.5σ above and below the mean, in such a distribution the range would be from about 1 to 99 (viz., $50 \pm 3.5 \times 14$). Scores transmuted into this, or any comparable distribution, are in reality standard scores, but are expressed in a somewhat more convenient form.†

Table 30 gives the raw scores made by Subject A upon the three tests of Table 29; A's standard scores; and his scores

TABLE 30

CONVERTING OBTAINED SCORES INTO EQUIVALENT SCORES IN A NEW DISTRIBUTION THE MEAN OF WHICH IS 50 AND THE σ 14

	Tests		
	Word-Building	Digit-Span	General Intelligence
Mean	16.3	7.4	150.4
σ	4.9	1.3	20.6
A's obtained scores	14	9	165
A's z -scores	-.47	+ 1.23	+ .71
A's converted scores	43	67	60

A's mean converted score = 57

converted into equivalent measures in a distribution whose mean is 50 and σ 14. The calculation of such converted scores is straightforward. Subject A's score in word-building is $-.47\sigma$ below the mean of the test distribution, and his equivalent score in the new distribution is $-.47 \times 14$ or -6.58 (i.e., 7) below the mean of 50. Subject A's score, therefore, is $50 - 7$ or 43. In the new distribution, 43 is as far below the mean of 50 as 14 is below the mean of 16.3 in the original score distribution. In the digit-span test, A's score is 1.23σ above the mean. His

* Hull, C. L., *The Conversion of Test Scores into Series which shall have any Assigned Mean and Degree of Dispersion*, Journal of Applied Psychology, 1922, 6, pp 298-300.

† The T -scores discussed in Chapter VI, page 151, are essentially standard scores in a distribution whose mean is 50 and whose σ is 10.

converted score, therefore, is 67; 17 units (1.23×14) above the mean of 50. In general intelligence, A's standard score of .71 is 10 units ($.71 \times 14$) above the mean of 50; and accordingly his converted score is 60 in the new distribution. A's average score on the three tests is 57.

Hull * has devised the following scheme by means of which scores in any distribution may be rapidly converted into equivalent scores in a normal distribution with a mean of 50 and σ of 14.

Let M = the mean of the given test

" σ = the σ " " " "

" X_1 = the individual's score in the given test

" 50 = the mean of the converted series

" 14 = the σ of the converted series

" X = the individual's score in the converted series.

Also, let $S = \frac{14}{\sigma}$

and $K = 50 - MS$;

then $X = K + SX_1$

To illustrate these formulas, in the word-building test A's score is 14, the mean of the test is 16.3, and the σ is 4.9. The mean of the new distribution is 50 and the σ is 14. Substituting these values in their appropriate equations, $S = 14/4.9$ or 2.9; $K = (50 - 16.3 \times 2.9)$ or 3, and

$$X = 3 + 2.9 \times 14$$

or

$$X = 3 + 40, \text{ or } 43$$

The advantage of the formula $X = K + SX_1$ is that once K and S have been calculated, equivalent scores in the new distribution may be found by simple substitution for X_1 in the equation. If Subject B, for instance, makes a score of 18 in the word-building test, his converted score will be

$$X = 3 + 2.9 \times 18 \text{ or } 55 \text{ (to nearest whole number).}$$

Converted scores in the other two tests in Table 30 have been calculated by the formulas given above. Scores which have

* *op. cit.*, p. 299.

been converted into a distribution in which the mean is 50, and which range from about 0 to 100, are easier to handle than *z*-scores. They are more intelligible, too, to persons without much statistical training. As pointed out above, converted scores are essentially standard scores. The assumptions made in comparing or combining equivalent scores are the same as the assumptions made in comparing or combining standard scores.

2. Converting Scores of Different Tests into Percentile Ranks

In percentile scaling, a child who makes a certain score upon a test is given a percentile rank of 27, 36, or 77, say, in accordance with his position in the distribution. When the distribution of each of several tests has been drawn up, individual scores may be readily translated into percentile ranks. These ranks may then be compared directly, or combined to give a final percentile ranking. The method of computing percentiles has already been considered (p. 76). It is only necessary here, therefore, to show how percentile rankings may be compared, or combined into a final score.

Table 31 gives the percentile distributions for nine year olds upon three tests of the Pintner-Paterson series of performance

TABLE 31
PERCENTILE DISTRIBUTIONS FOR 9 YEAR OLDS ON THREE TESTS.
METHOD OF COMBINING THE PERCENTILE RANKS
OF A SINGLE INDIVIDUAL

Tests	Percentiles											S's Score	S's Perc Rank
	0	10	20	30	40	50	60	70	80	90	100		
Picture Completion	62	240	297	325	372	407	440	450	499	577	646	445	65
Substitution	219	190	173	158	152	141	133	126	121	109	80	126	70
Seguin Form-Board	34	24	21	20	18	18	17	16	15	15	13	17	60
Median Percentile Rank.													65

tests.* The subject, a nine year old boy, made a score of 445 on the completion test which gives him a percentile rank of 65 (midway between 60 and 70). On the substitution test, a score of 126 gives him a percentile rank of 70; and on the Seguin form-board a score of 17 gives him a percentile rank of 60.

* Pintner, R., and Paterson, D. G., *A Scale of Performance Tests*, 1925, pp. 189 and 197.

The scores on tests 2 and 3 are in time units (seconds) so that the lowest score numerically represents the highest achievement.

The median of this subject's three percentile ranks is 65, which indicates that he stands somewhat above the median of children of his age in these tests. If this subject had been 10 or 11 years old, percentile distributions for these ages would, of course, have been used. Percentile ranks may be combined

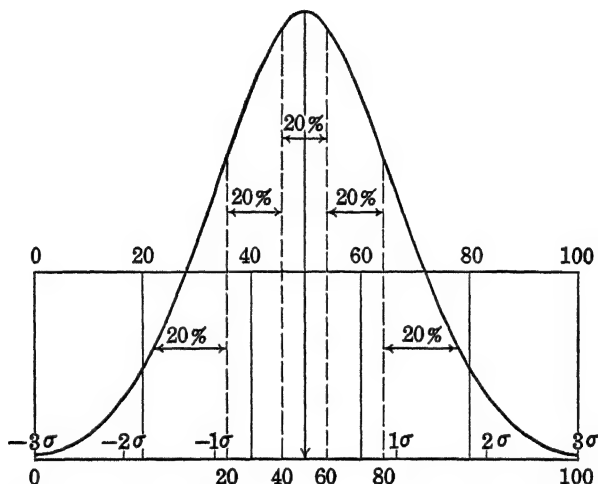


FIG. 36. To Illustrate the Position of the Same Five Percentiles in Rectangular and Normal Distributions.

directly, since such derived scores are expressed in comparable units and each test has equal weight in the final score.

Percentile scales assume that the difference between ranks of 10 and 20 is the same as the difference between ranks of 40 and 50; that is, percentile differences are taken to be equal throughout the scale. This assumption holds strictly, however, only when the distribution of scores is in the form of a rectangle rather than in the form of a normal curve. Figure 36 shows graphically the difference between the two types of distribution. The figure represents a rectangular distribution and a normal distribution of the same area plotted over it. The rectangular

distribution has been divided into five equal parts or quintiles by taking successive fifths of the area. Along the top of the rectangle, a linear scale composed of five equal units is laid off. The width of each small rectangle is the same — the distances from 0 to 20, from 20 to 40, from 40 to 60, from 60 to 80, and from 80 to 100 are all equal. Now let us compare these equal percentile distances with the same percentile distances calculated from the normal curve. The first 20% of area, counted off from the extreme left of the normal curve, covers almost twice the distance along the baseline of the curve as is occupied by the first 20% of the rectangular distribution. This first 20% also covers about four times as much of the baseline as the third 20% (i.e., that from 40 to 60) in the normal curve.* It is clear that the steps from 0 to 20 and from 20 to 40 are not equal when measured along the baseline of the normal curve. Note that this inequality is relatively greater at the extremes of the distribution than it is around the mean.

Since most distributions of test scores are normal or approximately normal, equal percentile distances cannot be taken to represent equal steps in difficulty throughout the percentile scale. Between Q_1 and Q_3 , however, percentile ranks are approximately equally spaced. Percentile ranks of a child in two different tests may be combined or averaged with little error when they fall between these limits. But percentile ranks greater than 75 or less than 25 should be combined, if at all, with full knowledge of their limitations.

3. The Median Mental Age Method

In transmuting test scores into age units, the mean number of points achieved on a test by unselected seven year olds is

* The baseline extent covered by the first 20% in the normal curve has been found in the following way. From Table 15, we find that the 30% of the area to the left of the mean extends from the mean to point $-.84\sigma$. Hence, the first 20% of the normal distribution falls between $-.84\sigma$ and -3.00σ . The second 20% lies between $-.84\sigma$ and $-.25\sigma$ since the last point lies at a distance of 10% from the mean. The third 20% lies between $-.25\sigma$ and $.25\sigma$. The fourth and fifth 20%'s occupy the same relative positions in the upper half of the curve as the second and first 20%'s occupy in the lower half of the curve.

scored 7; the mean number of points achieved by unselected eight year olds is scored 8, and so on for other age groups. An example of how test scores may be combined in terms of mental age units is given in Table 32. The mean scores on Tests *A*, *B*, *C*, *D*, and *E* for 7, 8, 9, 10, 11 and 12 year olds are tabulated in the body of the table.* Test *E* is scored in time units and hence the numerical score decreases with increases in age. Subject H's score on Test *A* is 25, and accord-

TABLE 32

TO ILLUSTRATE THE METHOD OF COMBINING SCORES WHICH HAVE BEEN CONVERTED INTO EQUIVALENT MENTAL AGES

Tests	Mean Score on Each Test at Ages:						H's Scores	H's Equivalent Mental Age Scores
	7	8	9	10	11	12		
A	10.2	14.0	18.3	25.0	30.6	37.5	25	10.0
B	75.0	79.2	84.3	86.5	90.1	96.3	88	10.4
C	5.2	7.3	9.1	12.4	16.3	21.2	10	9.3
D	110.2	115.0	121.3	127.3	134.5	138.8	136	11.4
E	230.7	224.2	210.0	203.6	192.5	180.0	210	9.0
H's Median Mental Age =								10.0

ingly his mental age score is 10 years. His score on Test *B* is 88, which falls between 86.5 and 90.1. By interpolation, 88, which lies $\frac{(88 - 86.5)}{(90.1 - 86.5)}$ or $\frac{1.5}{3.6}$ of one year (between age 10 and age 11), is equivalent to a mental age of 10.4. H's scores on Tests *C* and *D* are found by interpolation, in the same way as was his score on Test *B*, to be 9.3 and 11.4, respectively. H's *median mental age* on the five tests is 10 years.

The age scale has been widely used in psychological and educational measurement, and has the advantage of being easily interpreted. In scales of this sort the assumptions made are (1) that distributions of scores at each age level are approximately normal, and (2) that the age-progress "curves" found by plotting the mean test scores (on the *Y-axis*) against the

* In Table 32, age 7 means 6.5 to 7.5; that is, 7 is the midpoint of the interval. The same is true of the other ages.

ages (on the *X-axis*) are straight lines. The first assumption is valid for most mental tests. The second assumption is true up to the early 'teens, i.e., up to the time when marked negative acceleration usually occurs. Direct interpolation between age levels assumes straight line (or linear) relationship. But when the growth curve is no longer straight (bends in toward the *X-axis*) linear interpolation between successive age levels will not give consistent results.*

A practical disadvantage of the age scale is the difficulty which arises in getting unselected samples to be used in determining the norms of the low and the high age groups. Many very young children are not in school, while many of the brighter and duller older children have for various reasons been eliminated. Age scales are most valid when designed for use with children between the ages of 7 and 12.

4. Weighting Tests According to the Variability of Their Distributions

Suppose that several tests have been given, all by the time-limit or all by the amount-limit method, so that the resulting scores are all in *amount* or *time* units. If we wish to combine these tests into a composite or total score, or to average them, the separate tests should be weighted according to the variability of their scores and not according to the absolute sizes of the scores themselves. The illustration below will show why this is true. In the following table, Test 1 has more weight than Test 2 upon the composite score because the variability (range of scores) in Test 1 is greater than the variability (range

Subjects	Test 1	Test 2	Composite
A	30	165	195
B	45	170	215
C	60	175	235
Range:	<u>30</u>	<u>10</u>	<u>40</u>

* Garrett, H. E., and Schneek, M. R., *Psychological Tests, Methods and Results*, 1933, Part II, pp. 14-32.

of scores) in Test 2. The range of scores in the composite is 40. To this variability Test 1 contributes 30 and Test 2 10, that is, the two tests contribute to the variability of the composite in the ratio 3 : 1. Scores in Test 1, therefore, have three times the weight of scores in Test 2 in producing *differences* among the composite scores. If each score in Test 1 is divided by 3, or each score in Test 2 multiplied by 3, the variabilities (ranges) of the two tests will be the same; and the scores of the two tests will then enter with equal weight into the composite. In the table below scores on Test 1 have been divided by 3.

Subjects	"New" Test 1	Test 2	Composite
A	10	165	175
B	15	170	185
C	20	175	195
New Range:	$\overline{10}$	$\overline{10}$	$\overline{20}$

The ranges of the two tests are now equal, and each test contributes with equal weight to the range of 20 in the composite scores.

In weighting scores in accordance with their variability, σ or Q , and not the range, is the measure of variability ordinarily used. To illustrate weighting in terms of σ , suppose that in a given test in which the mean is 25 and σ is 5, Subject A scores 20; and that in another test in which the mean is 150 and σ is 25, A scores 160. If we add A's two scores ($20 + 160$) to get 180, the score in the second test is given 5 times as much weight in the composite as the score in the first, since the σ is five times as large in the second test. If it is desired to give the two tests equal weight, their σ 's must be equalized. This can be done by multiplying the σ of the first test by 5, or dividing the σ of the second test by 5. This same weighting procedure must then be applied to the separate scores. By the first operation our composite score becomes $20 \times 5 + 160$ or 260; by the second operation the composite score becomes $20 + \frac{160}{5}$ or 52. In either composite both tests will now have equal weight.

Table 33 illustrates the method of weighting the scores in

TABLE 33

TO ILLUSTRATE HOW SCORES ARE COMBINED WHEN WEIGHTED
ACCORDING TO THEIR VARIABILITY

(Data obtained from 200 Barnard Freshmen Women *)

	Tests				
	Logical Memory (recall)	Logical Memory (recognition)	Completion	Informa- tion	Vocab- ulary
Mean	6.50	37.47	35.78	104.71	73.90
σ	1.76	7.69	4.36	26.79	7.60
Multiplier to give all tests equal weight	5	1	2	$\frac{1}{3}$	1
New σ 's	8.80	7.69	8.72	8.93	7.60
A's scores	5	35	30	100	75
A's weighted scores (all tests equal)	25	35	60	33	75
A's weighted scores (tests 1 and 3 weighted 2, others 1)	50	35	120	33	75

several tests so that they may be combined with known weights. The scores of Subject A on each of five tests are given. If A's scores are added as they stand, Test 4 (information) will be given 15 times the weight of Test 1 (logical memory, recall), since the σ of the information scores (26.79) is 15 times the σ of the logical memory (recall) scores (1.76). Likewise, information will have approximately six times the importance of completion and approximately three times the importance of logical memory (recognition) and of vocabulary. It is hardly probable that information is as much superior as this to the other tests; in fact it is probably one of the least important in the battery. Hence, a fairer weighting seems desirable if the five scores are to be combined. The simplest plan at the start is to weight all of the tests equally as shown in the table. If we multiply the σ of Test 1 by 5, the σ of Test 2 by 1, the σ of Test 3 by 2, the σ of Test 4 by $\frac{1}{3}$, and the σ of Test 5 by 1, the σ 's are all approximately equal (see new σ 's in Table 33). Now if A's obtained scores are multiplied by these " σ -multipliers" his "derived"

* Carothers, F. E., *Psychological Examination of College Students*, Archives of Psychology, 1921, 46, pp 30-34.

test scores will all have the same weight in a final composite. In deciding upon multipliers, the best plan is to keep them whole numbers, if feasible, and as small as possible. In Table 33, the σ 's of Test 2 and 5 have been taken as "standards" because this permits simple multipliers for the other three tests.

Suppose it is desirable to give logical memory (recall) and completion *twice* as much weight as the other tests in the composite. To accomplish this, the σ 's of Tests 1 and 3 must be multiplied by 10 and 4 instead of by 5 and 2; their "new σ 's" must be twice as large as the σ 's of the other tests. In the present problem, since all of the tests have already been weighted 1, we need only double A's new scores on Tests 1 and 3. These tests now have twice the importance of Tests 2, 4 and 5 in the composite.

5. Converting Test Scores into Equivalent Ranks

The simplest method of comparing or combining scores from different tests is to rank the scores in each distribution in order of merit and then compare or average these ranks. The table below is an illustration of how the scores made by 10 subjects in two tests may thus be compared:

Subjects	Test 1 Scores	Test 2 Scores	Order of Merit Test 1	Order of Merit Test 2	Average Order of Merit Ranking	Final Average Order of Merit Re-ranking for Both Tests
A	20	65	8	3.5	5.75	6
B	28	52	2	10	6.00	7
C	19	63	9.5	5	7.25	9
D	22	71	5.5	1	3.25	1
E	31	58	1	7	4.00	2
F	26	60	3	6	4.50	4
G	24	53	4	8.5	6.25	8
H	22	65	5.5	3.5	4.50	4
I	21	67	7	2	4.50	4
J	19	53	9.5	8.5	9.00	10

The scores in Test 1 have been ranked in order of size from 31 (ranked first) down to 19. Since there are two scores of 19, instead of giving one a rank of 9 and the other a rank of 10,

each is ranked 9.5. The two scores of 22, which rank 5 and 6, are given average rankings of 5.5. The scores in Test 2 are ranked in order of size from 71 (highest) to 52 (lowest).

The relative position of each subject in the two tests may now be compared directly. Thus, Subject F ranks 3 on Test 1 and 6 on Test 2. If the order of merit rankings for the two tests are averaged, these averaged rankings may be re-ranked to give a measure of each subject's standing in the two tests. These re-rankings are given in the last column of the table. Note that each of the three average ratings of 4.5 (3, 4, and 5 in the order of merit series) has been given a median rank of 4.

The method of combining scores on different tests in terms of their ranks or relative positions is obviously a crude procedure. No account is taken of the size of the gaps in the series, of the variability within the series, or of the form of the distribution. Comparison in terms of ranks is most useful when the scores are too few or too scattered to justify the use of a more precise method.

II. THE MEAN AND SIGMA FROM COMBINED DISTRIBUTIONS

1. The Mean and σ When Two or More Distributions are Combined into a Single Distribution

Suppose that the mean score made by a group of 25 12 year old children on a learning test is 64, and that the mean score made by a second group of 200 children of the same age is 52. If these two groups are combined into a single distribution of 225 cases, what is the mean score of the whole group? When the number of children in two groups is the same, we need simply add the two means together and divide by two to obtain the mean of the combined group. But when, as in the present case, the number of subjects in the two distributions differ, we must weight each mean by the number of scores which it represents, add them, and divide by the number of cases in the combined distribution to obtain the mean of the combined group. Thus the mean of our combined distribution is

$\frac{25 \times 64 + 200 \times 52}{225}$ or 53.33. The formula for calculating the mean of a distribution composed of two distributions differing in size is

$$M = \frac{N_1 M_1 + N_2 M_2}{N} \quad (24)$$

(mean of a distribution composed of two component distributions)

in which M_1 and M_2 are the means of the two distributions; N_1 and N_2 are the numbers of cases in each distribution; and $N = N_1 + N_2$. Formula (24) may be readily extended to give the mean of a distribution composed of three, four or any number of groups.

If we have the σ 's of two different groups, we may calculate also the σ of the combined group, if we know the means of the two component distributions and the mean of the combined distribution. The method is as follows:

$$\begin{aligned} \text{Let } M_1 - M &= d_1 \\ M_2 - M &= d_2 \end{aligned}$$

in which M_1 is the mean of the first distribution and M_2 is the mean of the second distribution M is the mean of the combined distribution found from formula (24) above. Then

$$\sigma_{com} = \sqrt{\frac{N_1(\sigma_1^2 + d_1^2) + N_2(\sigma_2^2 + d_2^2)}{N}} \quad (25)$$

(σ of a distribution composed of two component distributions)

in which σ_1 and σ_2 are the σ 's of the two component distributions, N_1 and N_2 are the sizes of the component groups, N is the size of the combined group, and σ_{com} is the σ of the combined group.

We may illustrate formula (25) with a problem. The following data were obtained from two groups upon an educational achievement examination.

Group 1:	$M_1 = 61.43$	$\sigma_1 = 8.65$	$N_1 = 250$
Group 2:	$M_2 = 70.51$	$\sigma_2 = 7.39$	$N_2 = 90$

What is the σ of the combined distribution of 340 cases? From formula (24) we find that the mean of the combined distribution (M) is 63.83. Hence, d_1 , i.e. $(M_1 - M)$, is $61.43 - 63.83$ or -2.40 ; and d_2 , i.e. $(M_2 - M)$, is $70.51 - 63.83$ or 6.68 . Substituting these values in formula (25) we have

$$\begin{aligned}\sigma_{com}^2 &= \frac{250(74.82 + 5.76) + 90(54.61 + 44.62)}{340} \\ \sigma_{com}^2 &= 85.52 \\ \sigma_{com} &= 9.25\end{aligned}$$

Like formula (24), formula (25) may be easily extended (to give the σ of more than two combined distributions).

2. The Mean and σ of the Sum or the Difference of Two Series of Test Scores

If we know the correlation between two comparable series of test scores, X_1 and X_2 , and the σ 's of the two series, it is possible to compute the σ of the new composite series obtained by *adding* or *subtracting* the corresponding scores in the two original series. When the scores in the "new" distribution have been obtained by adding the corresponding scores, the formula for σ_s is

$$\sigma_s = \sqrt{\sigma_1^2 + \sigma_2^2 + 2r_{12}\sigma_1\sigma_2} \quad (26)$$

(σ of the sum of two series of scores, obtained by adding corresponding values)

in which σ_s is the σ of the new summed-series, σ_1 is the σ of the X_1 scores and σ_2 is the σ of the X_2 scores. The coefficient, r_{12} ,* denotes the correlation between the X_1 and X_2 scores. When the scores in the new distribution have been obtained by subtracting corresponding scores in the two tests, the formula is

$$\sigma_d = \sqrt{\sigma_1^2 + \sigma_2^2 - 2r_{12}\sigma_1\sigma_2} \quad (27)$$

(σ of the differences between two series of scores, obtained by subtracting corresponding values)

in which σ_d is the σ of the new difference-series.

* The calculation of r , the coefficient of correlation, is given in Chapter IX.

To illustrate, suppose that we have administered two controlled association tests, Verb-Object and Opposites. The mean of the first (X_1) is 42.00 with a σ_1 of 11.18; and the mean of the second (X_2) is 30.00 with a σ_2 of 9.00. The correlation between the two sets of scores, for the given group, is .60. The σ 's of the new series obtained by (1) adding corresponding X_1 and X_2 scores, and (2) subtracting corresponding X_1 and X_2 scores are

$$\begin{aligned}\sigma_s &= \sqrt{(11.18)^2 + (9.00)^2 + 2 \times .60 \times 11.18 \times 9.00} \\ &= 18.07\end{aligned}$$

and

$$\begin{aligned}\sigma_d &= \sqrt{(11.18)^2 + (9.00)^2 - 2 \times .60 \times 11.18 \times 9.00} \\ &= 9.23\end{aligned}$$

Thus, the σ of the summed-series is 18.07, and the σ of the differences is 9.23.

The mean of the summed-series is simply the sum of the two means divided by two. And the mean of the difference-series is the difference between the two means divided by two. In the first case, the mean is 36.00; in the second, 6.00.

PROBLEMS

- In a paired-associates test of memory, the mean score is 14.36 and the σ is 2.78. John's score is 16 and Henry's is 13.
 - Convert John's and Henry's obtained scores into z -scores.
 - Convert John's and Henry's obtained scores into equivalent scores in a distribution in which the mean is 50 and the σ 14.
 - Convert John's and Henry's obtained scores into equivalent scores in a distribution in which the mean is 100 and the σ is 20.
- Given the following five tests and scores made on each by John and James:

	Arithmetic	No. Series	Vocabulary	Reading	International Group Test (non-language)
M	40.39	18.84	39.26	44.14	193.98
σ	8.69	8.34	12.01	21.91	45.54
John	35	22	55	52	210
James	51	30	38	40	260

- (a) Convert John's and James's scores into z -scores. Average the five scores for each boy.
 - (b) Convert John's and James's scores into equivalent scores in a distribution the mean of which is 50 and the σ 10. Average the five converted scores for each boy.
 - (c) Combine John's scores and James's scores into a composite for each boy in which each score is weighted equally (viz., 1).
 - (d) Combine John's scores and James's scores into a composite in which arithmetic and reading have *twice* the weight of number series and vocabulary, and *three times* the weight of the International Test.
3. (a) Locate the deciles in a normal distribution in the following way. Beginning at -3σ , count off successive 10%'s of area up to $+3\sigma$. Tabulate the σ -values of the points which mark off the limits of each division. (For example, the limits of the first 10% from -3σ are -3.00σ and -1.28σ — see Table 15, p. 110.) Label these points in order from -3σ as .10, .20, etc. Now compare the distances in terms of σ between successive ten per cent points. Explain why these distances are unequal.
 - (b) Divide the baseline of the normal probability curve (take as 6σ) into ten equal parts, and erect a perpendicular at each point of division. Compute the percentage of total area comprised by each division. Are these percents of area equal? If not, explain why. Compare these percents with those found in (a).
 4. On the five tests given in Table 32, page 186, M's scores are as follows. Test A, 33; Test B, 88; Test C, 11; Test D, 125; and Test E, 190. Compute the median of the equivalent mental age scores.
 5. Five subjects score as follows upon three tests:

Subjects	Test 1	Test 2	Test 3
A	19	145	98
B	15	110	87
C	17	124	92
D	22	110	75
E	18	136	94

- (a) Rank the scores on each test in order of merit (put the largest as No. 1), average the ranks for each subject and re-rank to give a final order of merit.

- (b) Transmute the final rankings into units of amount (on a scale of 100) by means of Table 28, page 171.
6. Given the following data obtained from two groups upon the same test:

	Group 1	Group 2
N	100	900
M	146.32	138.65
σ	20.63	28.44

- (a) What is the mean of the distribution obtained by throwing the two groups together?
- (b) What is the σ of the distribution formed by combining the two groups?
7. The following data were obtained on two memory tests administered to the same group:

	Test 1	Test 2
M	15.6	20.2
σ	3.5	4.7

$$r_{12} = .75$$

- (a) Compute the mean and σ got by adding the corresponding scores in Tests 1 and 2.
- (b) Compute the mean and σ got by subtracting the corresponding scores in Tests 1 and 2.

ANSWERS

1. (a) John, .590; Henry, — .489
 (b) “ 58; “ 43
 (c) “ 112; “ 90
2. (a) John's z -scores in order: — .62; .38; 1.31; .36; .35. Mean = .36
 James's z -scores “ “ : 1.22; 1.34; — .10; — .19; 1.45.
 Mean = .74
- (b) John's converted scores: 44; 54; 63; 54; 54. Mean = 54
 James's “ “ : 62; 63; 49; 48; 65. Mean = 57
- (c) Using as σ -multipliers, 3, 3, 2, 1, $\frac{1}{2}$, we have
 John's scores: 105; 66; 110; 52; 105; composite, 438
 James's scores: 153; 90; 76; 40; 130; composite, 489

- (d) Multiply scores in (c) by 3, 1.5, 1.5, 3, 1, to get:
 John's scores. 315; 99; 165; 156; 105; composite, 840
 James's scores: 459; 135; 114; 120; 130; composite 958
3. (a) .00 .10 .20 .30 .40 .50 .60 .70 .80 .90 1.00
 - 3.00 - 1.28 - .84 - .52 - .25 0 .25 .52 .84 1.28 3.00
 Diffs: 1.72 .44 .32 .27 .25 .25 .27 .32 .44 1.72
- (b) Percents of area in order: .68; 2.77; 7.92; 15.92; 22.57;
 22.57; 15.92; 7.92; 2.77; .68.
4. 10.4 years
5. (a) Final order: *A*, *E*, *C*, *D*, and *B*.
 (b)

<i>A</i>	<i>E</i>	<i>C</i>	<i>D</i>	<i>B</i>
75	60	50	40	25
6. (a) 139.42
 (b) 27.85
7. (a) Mean = 17.9; $\sigma = 7.7$
 (b) Mean = 2.3; $\sigma = 3.1$

CHAPTER VIII

SAMPLING AND RELIABILITY

I. THE MEANING OF RELIABILITY

By the "true" measure of an individual's standing in any trait, as, for example, the true measure of his height, reaction-time, or intelligence, we mean the average of a very large (theoretically, "infinite") number of measurements of the given trait made under precisely the same conditions. In actual practice, one never deals with a true measure as thus defined; in fact, one must usually be satisfied with a single measure, or, at best, with comparatively few measures of an individual's ability. It is possible, however, to estimate the probable amount by which an individual's obtained score varies from its corresponding true score; and this estimate of "probable divergence" serves as an index of the *reliability* of the obtained score — of how good an approximation it is to the true score.

Like the reliability of an individual score, the reliability of an obtained measure of a *group* is also determined by finding the probable divergence of the obtained measure from the true measure of the group's capacity. The true measure of a group, as for example its true mean or true σ , is that hypothetical measure obtained by taking into account the entire population; that is, all of the members of the larger group from which the given group was drawn.* The true measure of the difference between two groups is the difference between their true means or medians.

To show more concretely just what is meant by the true measure of a group, let us suppose that we could measure the height of every 12 year old boy in the United States. If from this frequency distribution of heights, we should calculate a

* This group, of course, is usually a sample of a still larger group and so on. Hence, the true measure is never actually found.

measure of central tendency and a measure of variability, these would be "true measures." The mean and σ , for example, would represent the true mean height and the true variability of 12 year old boys in the United States. Moreover, if we could measure the height of every 12 year old girl in the United States, it would be possible to calculate the true mean height and the true σ of 12 year old girls in this country. Knowing the true mean height of 12 year old boys, and the true mean height of 12 year old girls, we should then be able to find the true *difference* between the mean heights of 12 year old boys and 12 year old girls in the United States

Unfortunately, it is rarely if ever possible to measure *all* of the individuals in a given population; and it is, of course, clearly impossible to take an infinite number of measures of a single person. We must be content, therefore, to deal with "samples" selected from the total number of possible measures; and as a result, owing to slight differences in the samples drawn, measures of central tendency and of variability will usually be somewhat larger or somewhat smaller than their corresponding true measures. Whenever we have measured an individual or a group, therefore, we must ask ourselves this question: "How reliable a measure of capacity have I obtained? How well does it represent the true measure which I should get from a very large (infinite) number of measures of this individual, or from measuring *all* of the individuals in the population from which my group was taken?" This question will often lead to a second: "How many measurements must I make in order to obtain a result which shall meet a given standard of reliability, i.e., have a probable divergence from the true result which is less than some given amount?"

The purpose of this chapter is to develop methods which will enable us to answer these questions. The reliability of the measures of central tendency will be first considered; then the reliability of measures of variability and of certain other measures; and finally the reliability of the difference between measures.

II. THE RELIABILITY OF MEASURES OF CENTRAL TENDENCY

A. The Reliability of the Mean

(1) The Standard Error of the Mean (σ_M)

Perhaps the simplest approach to the study of the reliability of the mean is to examine the factors upon which the stability of this measure must depend. Suppose that we wish to find the mean score of college freshmen in the United States upon the American Council Psychological Examination. To measure the achievement of college freshmen in general would require in strict logic that we test *all* of the freshmen in the United States. But as this is obviously a well-nigh impossible task, we must be satisfied with taking the records of as *large* and as *random* a sample of freshmen as we can find. This means that we cannot use freshmen from only a single institution or from only one section of the country; and that we must guard against selecting only those with high, or only those with low, scholastic records. The more successful we are in getting an "unselected" group, the more nearly representative will this group be of all of the freshmen in the country. Evidently, therefore, the reliability (the "representativeness") of a mean depends for one thing on how impartially we have chosen our sample.

Given a fair sample, the reliability of a mean depends mathematically upon two characteristics of the distribution,* (1) the number of cases (N) and (2) the variability or spread of the measures in the sample.

(1) It is clear that the number of cases must influence the stability of a mean, since the addition of even one extra measure to a series will change the mean unless the additional case happens to coincide with the mean exactly. Moreover, the addition of one case to a set of 10 measures will bring about a greater change in the obtained mean than the addition of one case to a set of 1000 measures, as each case counts for less in the larger group. It may be shown mathematically, as well

* The reliability of a mean also depends upon the errors of measurement in the scores themselves (p. 332).

as experimentally,* that the reliability of an obtained mean will increase, not in proportion to the number of measures upon which it is based, but in proportion to the square root of the number of measures. The mean obtained from 25 measures, for example, is not 25 times, but $\sqrt{25}$ or 5 times as reliable as a single measure. And a mean based upon 36 cases is not 4 times as reliable as a mean based upon 9 cases, but only twice as reliable — since $\sqrt{36}$ divided by $\sqrt{9}$ equals 2.

(2) In addition to the size of the sample, the reliability of a mean also depends upon the variability of the separate measures around the mean. If the σ of the distribution is large, the separate measures tend to scatter widely around the mean, and we are unable to say where those cases in the population which we have not measured will most probably fall — whether they will be close to, or far from, the obtained mean. On the other hand, if the σ is small, we may be fairly certain that unmeasured cases will fall close to the mean. The reliability of an obtained mean, therefore, depends upon the size of the σ ; as σ increases, the reliability decreases.

To summarize the above discussion, the reliability of a mean depends *first* upon our having drawn a representative sample from the larger group or population which we are studying. When this condition has been met, and only then, the reliability of a mean is measured mathematically by its standard error which is based upon N (the number of cases) and the σ of the distribution. The formula for the standard error of the mean is

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \quad (28) \approx$$

(the standard error of the arithmetic mean) †

* Yule, G. U., *An Introduction to the Theory of Statistics*, 9th Edition, 1929, p. 257. For results of experiment, see Thorndike, E. L., *Empirical Studies in the Theory of Measurement*, Archives of Psychology, 1907, 3, pp. 1-13.

† When N is small (less than 30, say) formula (28) will give more accurate results when written

$$\sigma_M = \frac{\sigma}{\sqrt{N-1}}.$$

This is one of the most important, and most often used, of the reliability formulas. The standard error of the mean measures the extent to which the mean is affected by errors of measurement (p. 332) as well as by fluctuations which arise from sampling. A decrease in σ or an increase in N will cause the standard error to become smaller numerically. A decrease in σ_M means that the probable divergence of the obtained mean from the true mean of the population is just so much less. Hence, the reliability of an obtained mean increases as σ_M decreases.

A problem will illustrate the use and interpretation of formula (28).

*Problem (1) ** In 1883, the Anthropometric Committee of the British Association found the mean height of 8585 adult males in the British Isles to be 67.46 inches, with a σ of 2.57 inches. How reliable is this mean? By how much does it probably diverge from the mean which would have been obtained had all adult males in the British Isles been measured?

Applying formula (28) we find the standard error of the mean, σ_M , to be $\frac{2.57}{\sqrt{8585}}$ or $\underline{.028}$ inch. This result may be interpreted in the following way. The chances are about 68 in 100 that the obtained mean of 67.46 inches does not diverge from the true mean by more than $\pm .028$ inch, that is, by more than $\pm 1\sigma_M$ (Table 15). Stated in another way, the chances are 68 in 100 that the true mean lies within the limits $67.46 - .028$ and $67.46 + .028$ or between 67.43 and 67.49 inches. We may be practically certain that the true mean of the population lies within the limits $67.46 \pm 3 \times .028$ ($\pm 3\sigma_M$), or between 67.38 and 67.54 inches.

How the standard error measures the reliability or stability of the obtained mean of 67.46 inches may be still more clearly shown perhaps in the following way. Suppose that we have

* Yule, G. U., *An Introduction to the Theory of Statistics*, 9th Edition, 1929, pp. 112 and 141.

calculated the mean height of each of 1000 groups of men; that each group contains 8585 subjects, and that the groups or samples are drawn at random from the general population. The 1000 means obtained from these groups will tend to differ slightly from one another due to sampling fluctuations (p. 245) and to errors of measurement (p. 314). Hence, not all samples will represent with equal fidelity the population from which they have been drawn. Now assume that it were possible to obtain the mean height of the *entire male population* of the British Isles. If we should subtract this *true* mean height from each of the 1000 obtained mean heights, 1000 differences would be obtained. The best assumption we can make is that the frequency distribution of these 1000 differences will follow the normal probability curve with a mean at zero (p. 107). In this hypothetical normal distribution-of-differences, we should have relatively few *large* plus or minus differences; and many *small* plus, *small* minus, and *zero* differences. In short, the obtained means will hit very near to the true mean, or fairly close to it, more often than they will miss it by large amounts.

The central tendency of our distribution of 1000 differences will fall at zero, since zero will be the difference most often obtained if our samples are "randomly drawn" from their population. The σ of this distribution-of-differences is the standard error of the mean, σ_M . In other words, formula (28) measures the spread of the "obtained mean-true mean differences" around zero as a measure of central tendency. It is because of this fact that the standard error of the mean becomes a measure of the amount by which the obtained mean probably diverges from the true mean.

The results of our hypothetical experiment are represented graphically in Figure 37, page 204. The 1000 differences between the 1000 obtained means and the true mean are represented by a normal frequency distribution with mean at zero and σ equal to .028. The heights of the different ordinates (y 's) represent the frequency of the various (obtained mean-true mean) differences. That zero is the most frequently obtained differ-

ence is shown by the fact that the ordinate at the mean is the maximum ordinate. We have learned that the σ of a normal distribution includes the middle 68.26% of the cases when measured off in the plus and minus direction from the mean. Hence we know that the chances are 68 in 100 that the difference between the obtained mean of 67.46 inches and the true

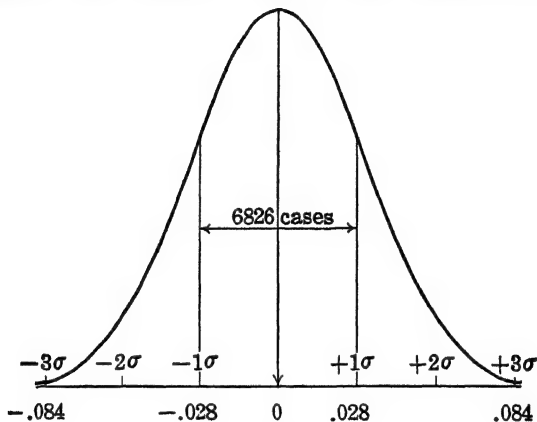


FIG. 37.

mean will not be greater than $\pm .028$ inch. Or, as stated above, there are 68 chances in 100 that the true mean lies within the limits 67.49 and 67.43 inches.

The mean height of our sample of 8585 British males was given as 67.46 inches; and the standard error of this mean has been found to be .028 inch. From these data, let us try to answer the second question proposed on page 199, namely, "How many measurements must I make in order to get a result whose probable divergence from the true result is *less* than some given amount?" Specifically stated, suppose we wish the mean of 67.46 inches to be *twice* as reliable as it is — how many cases will be required? If the obtained σ of 2.57 inches remains approximately equal to the given σ when the group is increased in size, all that we need to do in order to cut σ_M , the standard error, in two, and thus double the reliability, is to place a "2"

in the denominator of the fraction $\frac{2.57}{\sqrt{8585}}$. But $2 \times \sqrt{8585}$ becomes $\sqrt{4 \times 8585}$ when the 2 is placed under the radical; and hence it is evident that 8585 must be multiplied by 4 in order for σ_M to be one-half its original size. If we must multiply N by 4 in order to *double* the reliability of a mean, we must multiply N by 9 in order to *triple* the reliability; by 16 in order to *quadruple* the reliability, and so on. If the σ of the distribution remains substantially constant as N is increased, the mean obtained from 400 cases is twice as reliable as the mean obtained from 100 cases; and the mean obtained from 900 cases three times as reliable as the mean obtained from 100 cases.

(2) The PE of the Mean (PE_M)

The reliability of a mean may be determined by PE_M instead of by σ_M . PE_M may be obtained by multiplying σ_M (i.e., $\frac{\sigma}{\sqrt{N}}$) by .6745 (see p. 114). Thus

$$PE_M = \frac{.6745\sigma}{\sqrt{N}} \quad (29)$$

(the probable error of the arithmetic mean) *

PE_M is interpreted in the same way as σ_M . In the problem of heights described on page 202, PE_M is found to equal .019 inch. The chances are *even*, therefore, that the obtained mean of 67.46 inches does not differ from the true mean by more than $\pm .019$ inch. Furthermore, since $\pm 4PE$ includes practically all of the cases in the normal distribution, we can be fairly certain (the chances are 99 in 100) that the true mean lies within the limits $67.46 \pm 4 \times .019$ or between 67.38 and 67.54 inches (see Table 16).

A comparison of the extreme limits within which we may be practically certain that the true mean lies shows that these limits usually differ slightly when $\pm 4PE_M$, instead of $\pm 3\sigma_M$, are taken as limiting points. This is true because $\pm 4PE$ in-

* This formula is written $PE_M = \frac{.6745\sigma}{\sqrt{N-1}}$ when N is less than 30.

cludes a slightly smaller percentage of the total distribution than $\pm 3\sigma$ (see Tables 15 and 16). The $\pm 3\sigma$ limits include 43 more cases than the $\pm 4PE$ limits, and while 43 cases in 10,000 may seem to be an insignificant number (and is relatively insignificant if taken from the middle of the distribution) even so few cases as 43 will have considerable importance at the extremes of the distribution. This may be seen from the fact that we must take $\pm 4.45PE$ in order to include *exactly* the same number of cases in a normal distribution as are included by $\pm 3\sigma$.

It is customary, however, in measuring reliability, to take $\pm 4PE$ instead of $\pm 4.45PE$ as the limits of virtual certainty. In the first place, $\pm 4PE$ mark off limits within which the chances are very great (9930 in 10,000) that the true mean will fall. Furthermore, the slightly closer approach to certainty got by using $\pm 4.45PE$ instead of $\pm 4PE$ is usually not sufficient to offset the greater convenience of the latter figure.

2. The Reliability of the Median

When the frequency distribution is normal or closely normal, the reliability of an obtained median may be calculated by making slight changes in the formulas for finding the reliability of the mean. The σ_{Mdn} and PE_{Mdn} are 1.2533 (roughly $\frac{5}{4}$) times the σ_M and the PE_M , respectively. Thus

$$\sigma_{Mdn} = \frac{1.2533\sigma}{\sqrt{N}} \quad (30)$$

(standard error of the median for normal distributions)

$$PE_{Mdn} = \frac{1.2533 \times .6745\sigma}{\sqrt{N}} = \frac{.8454\sigma}{\sqrt{N}} \quad (31)$$

$$\text{or} \quad PE_{Mdn} = \frac{1.2533Q}{\sqrt{N}} \quad (32)$$

(probable error of the median for normal distributions)

Formulas (30), (31) and (32) are applied and interpreted in the same way as are the standard and probable error formulas of the mean. A problem will illustrate their use.

Problem (2) On the Trabue Language Scale A,* 801 twelve year old boys made the following record: Median = 21.4; $Q = 4.9$. Assuming the distribution to be normal, how reliable is this median? How well does it represent the median of twelve year old boys in general on the given scale?

From formula (32), PE_{Median} is found to be .22. The chances are 50 in 100, therefore, that the true median does not differ from 21.4 by more than $\pm .22$. We may be reasonably assured that the true median lies within the limits $21.4 \pm 4 \times .22$ or between 20.5 and 22.3.

Both σ_{Median} and PE_{Median} are larger by approximately 25% than the corresponding measures of reliability of the mean. Hence, when the distribution is normal, the mean is a more reliable measure of central tendency than is the median. But this is not always true, for if the distributions are significantly 'peaked' (p. 117) the median may be the more reliable measure. A general formula for calculating the standard error of the median, when one has the entire frequency distribution, is

$$\sigma_{\text{Median}} = \frac{i\sqrt{N}}{2F} \quad (33)$$

(standard error of the median, normality not assumed)

where i equals the step-interval, F equals the frequency on the step which contains the median, and N is the number of cases in the sample.

When the frequency distribution is substantially normal, formulas (30) and (33) give essentially the same results. But for leptokurtic distributions and those containing a few very large or very small measures formula (33) is often the better measure. One source of instability in formula (33) should be pointed out. This is the fact that F (frequency on the step containing the median) depends upon the step-interval used in grouping, and hence may vary considerably for different classifications.

* Trabue, M. R., *Completion Test Language Scales*, Teachers College, Columbia University, Contributions to Education, 1916, 77, p. 15.

Kelley * recommends "smoothing" the frequency in the neighborhood of the median so as to get a closer approximation to the "theoretically correct" F . A simple test for determining whether the median or the mean is the more reliable measure of central tendency for a given distribution has been suggested by Shen.† This test consists in calculating the ratio $\frac{\sigma_{Mdn}}{\sigma_M}$ which equals $\frac{iN}{2F\sigma}$. If $\frac{\sigma_{Mdn}}{\sigma_M} > 1$, the M will, in general, be the more reliable measure; if $\frac{\sigma_{Mdn}}{\sigma_M} < 1$, the Mdn will be the more reliable measure.

III. THE RELIABILITY OF MEASURES OF VARIABILITY

1. The Reliability of the Standard Deviation, or σ

We have learned in the preceding section that the reliability of an obtained mean or of an obtained median is found by calculating the probable discrepancy between the obtained measure and its theoretically true value. In exactly the same way, the reliability of an obtained σ is determined by calculating the probable discrepancy between the obtained σ (i.e., the σ of the sample) and the true σ . The true σ is the σ which one would get from the population from which our sample was drawn. The formula for calculating the reliability of an obtained σ is

$$\sigma_{\sigma} = \frac{\sigma}{\sqrt{2N}} \quad (34)$$

(standard error of a standard deviation) ‡

In the problem on page 202, we found that for 8585 adult British males the σ taken around the mean of 67.46 inches was 2.57

* Kelley, T. L., *Statistical Method*, 1923, pp. 90-91.

† Shen, E., *Note on the Sampling Error of the Median*, *Journal of Educational Psychology*, 1935, 26, pp. 154-156.

‡ When N is less than 30, formula (34) should be written

$$\sigma_{\sigma} = \frac{\sigma}{\sqrt{2(N-1)}}$$

inches. How reliable is this σ ? How well does it represent the true σ which we should get if deviations were taken from the true mean height? Substituting for σ and for N in formula (34), we find the value of σ_σ to be .020 inch. This means that the chances are 68 in 100 that 2.57 inches does not differ from the true σ by more than $\pm .020$ inch. Further, the chances are 997 in 1000 that the obtained σ does not differ from the true σ by more than $3 \times .020$ or .060 inch. We may be almost certain, therefore, that the true σ lies within the limits $2.57 \pm .060$ or between 2.51 and 2.63 inches.

2. The Reliability of the Quartile Deviation, or Q

The reliability of the Q of a distribution may be found from the formula

$$\sigma_Q = \frac{1.11\sigma}{\sqrt{2N}} \quad (35)$$

*(standard error of Q in terms of the σ of the distribution) **

or from the formula

$$\sigma_Q = \frac{1.65 Q}{\sqrt{2N}} \quad (36)$$

*(standard error of Q in terms of the Q of the distribution) **

On page 207, the median score of the 801 twelve year old boys who took the Trabue Completion Test, Scale A, was 21.4 with a Q of 4.9. What is the reliability of this Q ? From formula (36) σ_Q is found to be .20. The chances are 68 in 100, therefore, that 4.9, the obtained Q , does not diverge from the true Q by more than $\pm .20$. And the chances are 997 in 1000 that the true Q lies within the limits $4.9 \pm 3 \times .20$ or between 4.3 and 5.5.

* When N is less than 30, formulas (35) and (36) should be written

$$\sigma_Q = \frac{1.11\sigma}{\sqrt{2(N-1)}} \quad \text{and} \quad \sigma_Q = \frac{1.65 Q}{\sqrt{2(N-1)}}.$$

IV. THE RELIABILITY OF THE DIFFERENCE BETWEEN TWO MEASURES

1. The Reliability of the Difference Between Two Means(1) The Standard Error of the Difference when Means are Uncorrelated (σ_D)

Suppose that we wish to find whether there is any difference between ten year old boys and ten year old girls in their knowledge of words. The usual method of attacking such a problem as this is to select a large and random sample of ten year old boys and ten year old girls; administer a vocabulary test; compute the mean scores; and find the difference between the two means. If this difference is five points, let us say, in favor of the girls, such a result — on the face of it — is evidence for believing that the typical girl knows more words than the typical boy. Even this tentative conclusion, however, is not strictly warranted, if all that we have is the obtained difference. And before we can draw a *definite* conclusion, we must know how closely the obtained difference of five points approximates to the true difference between ten year old boys and girls, i.e., the difference which we should get if we could compare the true mean score of the girls with the true mean score of the boys. Unless the obtained difference between the mean scores of boys and girls is *significant*,* it is entirely possible that if we compared the mean scores of other groups of boys and girls similarly selected as our own groups, the difference obtained might be zero, or even reversed in favor of the boys.

Clearly, then, it is important that we have some way of estimating the reliability of an obtained difference; that is, some way of telling whether one group is *sufficiently* superior to another to enable us to say with confidence that no matter how often similar groups are compared, the first group will nearly *always* excel the second. Furthermore, and equally important, if the obtained difference is *not* significant, we

* An obtained difference is significant when the odds are great that the true difference is greater than zero.

should know, if possible, how near it approaches to significance.

The formula for calculating the significance of the difference between two obtained means, when we are dealing with different groups, is

$$\sigma_D \text{ OR } \sigma_{M_1-M_2} = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2} \quad (37)$$

(standard error of the difference between two
uncorrelated means) *

in which σ_{M_1} is the standard error of the mean of the first group; σ_{M_2} is the standard error of the mean of the second group; and σ_D is the standard error of the difference between the two means. From formula (37) it is clear that, to find the reliability of the difference between two means, we must first know the reliability of the means themselves.

The application and interpretation of formula (37) may be illustrated by the following problem:

Problem (1) In a study of the intelligence of the foreign-born white draft during the World War, a sample of 308 native-born Germans and a sample of 325 native-born Danes were found to test as follows upon the "combined scale." †

Country of Birth	Number of Cases	Mean Score	σ
Germany	308	13.88	2.43
Denmark	325	13.69	2.23

The difference between the two obtained means is .19 (13.88 - 13.69) in favor of the Germans. Is this difference significant? Would further testing of similar groups of Germans and Danes give virtually the same results; or is it probable that the difference would be reduced to zero, or even reversed in favor of the Danes? To answer these questions we must first calculate the reliability of the mean scores of Germans and Danes,

* Means are uncorrelated when calculated from different groups, or from uncorrelated tests given to the same group

† The "combined scale" included the eight Alpha tests, the Stanford Binet, and tests 4, 5, 6, and 7 from Beta. The maximum score was 25. For the data given in this problem, see Brigham, C. C., *A Study of American Intelligence*, 1923, pp. 120-121.

and from these the reliability of the difference between the means. By formula (28), the standard errors of the two means are

$$\text{Germans:} \quad \sigma_{M_1} = \frac{2.43}{\sqrt{308}} = .138$$

$$\text{Danes:} \quad \sigma_{M_2} = \frac{2.23}{\sqrt{325}} = .124$$

Substituting these standard errors in formula (37), we have

$$\begin{aligned} \sigma_D &= \sqrt{(.138)^2 + (.124)^2} \\ &= .186 \end{aligned}$$

The actual difference between the two means is .19, and σ_D , the standard error of this difference, is .186. The obtained difference between the two means is interpreted in terms of its standard error in exactly the same way as a mean is interpreted in terms of its standard error. Thus, we say that the chances are 68 in 100 that the obtained difference of .19 does not differ from the true difference by more than $\pm .186$ (i.e., $\pm .19$); and that the chances are 99 in 100 that the obtained difference of .19 does not differ from the true difference by more than $\pm 3 \times .186$ or $\pm .56$.

We may be practically certain, then, that the true difference between the means of Germans and Danes lies within the limits $-.37$ and $+.75$ ($.19 \pm .56$). Since the lower limit of this range is negative, the obtained difference is clearly *not* significant. There is *some* chance, in other words, that the true difference is *less* than zero — i.e., that the true mean of the Germans is less than the true mean of the Danes. What are the chances that the Germans will, on the average, always be superior to the Danes on these tests? The chances that any obtained difference is significant may be read conveniently from Table 34. To illustrate, when D is .19, and σ_D is .186, so that D/σ_D is 1.0, the chances are 84 in 100 that the true difference is greater than zero. There are, therefore, 84 chances in 100 that the Germans will, on the average, *always* score above the Danes; or, put

TABLE 34

TO FIND THE CHANCES OF A SIGNIFICANT DIFFERENCE, I.E., TO FIND THE CHANCES THAT THE TRUE DIFFERENCE IS GREATER THAN ZERO, GIVEN THE OBTAINED DIFFERENCE BETWEEN TWO MEASURES, AND THE STANDARD ERROR OF THE DIFFERENCE

Example: A D/σ_D of 1.3 means that the chances are 90 in 100 that the obtained difference is significant — that the true difference is greater than zero.

$\frac{D}{\sigma_D}$	Chances in 100	$\frac{D}{\sigma_D}$	Chances in 100
.00	50	1.15	87
.05	52	1.20	88
.10	54	1.25	89
.15	56	1.30	90
.20	58	1.35	91
.25	60	1.40	92
.30	62	1.45	93
.35	64	1.50	93
.40	65	1.60	94
.45	67	1.70	96
.50	69	1.80	96
.55	71	1.90	97
.60	73	2.00	98
.65	74	2.10	98
.70	76	2.20	99(98.6)
.75	77	2.30	99(98.9)
.80	79	2.40	99(99.2)
.85	80	2.50	99(99.4)
.90	82	2.60	99(99.5)
.95	83	2.70	100(99.7)
1.00	84	2.80	100(99.74)
1.05	85	2.90	100(99.8)
1.10	86	3.00	100(99.9)

differently, that the true difference between the mean scores of the two groups is somewhat greater than zero, and is in /- favor of the Germans.

It is customary to take a D/σ_D of 3 as indicative of a significant difference (virtual certainty) since there is only about 1 chance in 1000 that a difference of $+3\sigma$ will arise when the true difference is zero.* In the problem above, the D/σ_D of

* It will aid, perhaps, in understanding better just what is meant by the reliability of the difference between means if we consider the following hypothetical situation. Suppose that we could test 1000 groups of native-born Germans and 1000 groups of native-born Danes upon the "combined scale." Assume that our samples are drawn at random from the general

TABLE 35

TO FIND THE CHANCES OF A SIGNIFICANT DIFFERENCE, GIVEN
THE OBTAINED DIFFERENCE BETWEEN TWO MEASURES AND
THE PROBABLE ERROR OF THAT DIFFERENCE

Example: A D/PE_D of 1.10 means that there are 77 chances in 100 that the obtained difference is significant, namely, that the true difference is greater than zero.

$\frac{D}{PE_D}$	Chances in 100	$\frac{D}{PE_D}$	Chances in 100
.00	50	1.55	85
.05	51	1.60	86
.10	53	1.65	87
.15	54	1.70	87
.20	55	1.75	88
.25	57	1.80	89
.30	58	1.85	89
.35	59	1.90	90
.40	61	1.95	91
.45	62	2.00	91
.50	63	2.10	92
.55	64	2.20	93
.60	66	2.30	94
.65	67	2.40	95
.70	68	2.50	95
.75	69	2.60	96
.80	71	2.70	97(96.6)
.85	72	2.80	97
.90	73	2.90	97(97.5)
.95	74	3.00	98(97.9)
1.00	75	3.10	98
1.05	76	3.20	98(98.5)
1.10	77	3.30	99(98.7)
1.15	78	3.40	99(98.9)
1.20	79	3.50	99
1.25	80	3.60	99
1.30	81	3.70	99
1.35	82	3.80	99(99.5)
1.40	83	3.90	100(99.6)
1.45	84	4.00	100(99.7)
1.50	84		

population of native-born Germans and native-born Danes and are roughly, at least, of the same size as the samples we have (i.e., about 300). If these 1000 groups were paired off randomly group against group, we should have 1000 differences between the means of Germans and Danes, each difference being analogous to the actually obtained difference of .19.

Now what we really want to know is the probability that the true difference between the mean scores of Germans and Danes is greater than zero. To test this hypothesis, we assume that our "distribution-of-differences" follows the normal probability curve, and construct a normal

1.0 is $\frac{1}{3}$ of what it should be, namely 3.0, to insure a significant difference; i.e., to guarantee that the true difference is greater than zero. A D/σ_D greater than 3.0 may be taken as indicating just so much additional security.

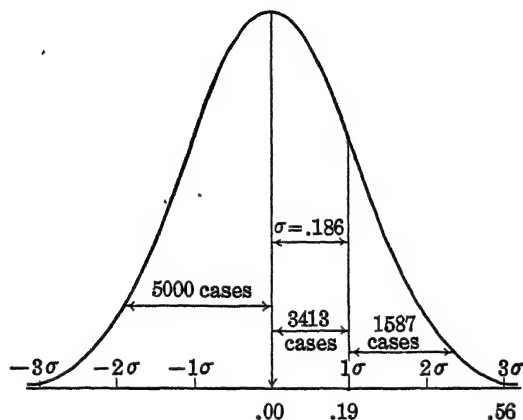


Fig. 38.

(2) The Probable Error of a Difference When Means are Uncorrelated (PE_D)

The reliability of the difference between means obtained from two different groups may be estimated by the PE_D instead of the σ_D . The formula for PE_D is

$$PE_D = \sqrt{PE_{M_1}^2 + PE_{M_2}^2} \quad (38)$$

(probable error of the difference between two uncorrelated means)

curve (see Fig. 38) with a mean at zero and a σ of .186. (Our best estimate of the σ of this distribution of 1000 differences is the obtained σ_D .) If the true mean of our 1000 differences were actually zero, and the σ_D were .186, it is clear that *some* of these 1000 differences would be as large as .19, and that some would be still larger (see Fig. 38). In that part of our curve-of-differences between .00 and .19 (mean + 1σ) lie 34.13% of the differences (Table 15). Above .19 (the obtained difference) are 15.87% of the differences, and below .19 are 34.13%, *plus* the 50.00% in the lower half of the curve. Of our 1000 differences, therefore, 84% would be expected to be *less* than .19, and 16% equal to or greater than .19, when the true difference is zero. Thus, a difference as large as .19 (or larger) could occur only 16% of the time — i.e., 160 times in our 1000 differences, if the true

in which PE_{M_1} and PE_{M_2} are the PE 's of the given means. Formula (38) is interpreted in exactly the same way as formula (37). A problem will illustrate its use.

Problem (2) On an arithmetic reasoning test,* 189 ten year old boys, and 206 ten year old girls made the following scores.


	Mean	σ
Boys	40.39	8.69
Girls	35.81	8.33

Is the difference between the means of the boys and girls significant, that is, large enough to guarantee that the true difference between the mean abilities of the two groups in solving arithmetic problems is greater than zero? First, we must calculate the PE 's of the two means:

$$\text{Boys: } PE_{M_1} = \frac{.6745 \times 8.69}{\sqrt{189}} = .426 \text{ by formula (29)}$$

$$\text{Girls: } PE_{M_2} = \frac{.6745 \times 8.33}{\sqrt{206}} = .392 \text{ by formula (29)}$$

difference is zero (as we have supposed). But, .19 *did* actually occur, and hence the chances are 16 in 100 that the true difference is zero (or negative), and 84 in 100 that the true difference is *somewhat greater* than zero. Stated in another way, we may say that the odds are 84 to 16 or about 5 to 1 *against* our original hypothesis — namely, that the true difference is, in fact, zero.

It seems clear that the obtained difference of .19 is large enough to insure considerably more than an even chance of a significant difference between the means of Germans and Danes upon the "combined scale." But it is not large enough to *guarantee* that the mean score of the Germans will *always* be higher than the mean score of the Danes. The further question arises, therefore, of how much difference is required to insure significance — to guarantee that the mean score of the Germans will always be higher than that of the Danes. The answer to this question can readily be given with the help of Figure 38. If the obtained difference were .56 (i.e., at $+3\sigma$ in Fig. 38) there would be only 1 chance in 1000 (Table 15) that such a difference could have arisen *if* the true difference were zero. Expressed in another way, if the difference of .56 had been obtained, there would have been 999 chances in 1000 that the true difference is greater than zero; or the odds would have been 999 to 1 that our hypothesis of a true difference of zero is untenable. 

* Schiller, B., *Verbal, Numerical, and Spatial Abilities of Young Children*, Archives of Psychology, 1934, 161, p. 21.

Substituting PE_{M_1} and PE_{M_2} in formula (38), we have

$$\begin{aligned} PE_D &= \sqrt{(.426)^2 + (.392)^2} \\ &= .579 \end{aligned}$$

The obtained difference is 4.58 and the PE_D is .579. The D/PE_D is 7.91. From Table 35 we find that a D/PE_D of 4 indicates complete reliability, i.e., is significant. It follows, therefore, that our obtained difference is not only significant, but is 3.91 (7.91 - 4.00) or almost twice as large as it needs to be in order to guarantee that the true difference is greater than zero.

The ratio D/PE_D is sometimes called the "critical ratio," because it provides a way of telling whether one group is significantly superior, on the average, to another in performing a given task. Just as it is customary to take a D/σ_D of 3 as evidence of significant superiority, so D/PE_D must be at least 4 in order to insure significant superiority in the direction indicated by the obtained result.

(3) The Standard and Probable Errors of the Difference Between Two Means, When the Means are Correlated

The last sections have dealt with the problem of discovering whether the difference between two means is significant when the means represent the performances of different groups—boys and girls, Germans and Danes, and the like (see also p. 216). A somewhat different, but closely related, problem is concerned with the reliability of the difference between two means obtained from the same group on the same test administered upon different occasions.* Suppose, for example, that we have administered a certain test to a group of children, and after two weeks have repeated the test. We wish to measure the effect of practice, or of intervening training, upon the final score; or to estimate the effect of some activity interpolated between testing and retesting. In order to find the

* This method is often called the "single group" method.

significance of the difference between the two means in the final and initial testing, we must use the formula

$$\sigma_D = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2 - 2r_{12}\sigma_{M_1}\sigma_{M_2}} \quad (39)$$

(standard error of the difference between correlated means)

in which σ_{M_1} and σ_{M_2} are the standard errors of the initial and final test means, and r_{12} * is the coefficient of correlation between the scores made on the initial and final tests.† An illustration will bring out the difference between formula (37) and formula (39).

Problem (8) At the beginning of the school year, the mean score of a group of 65 sixth grade children upon an educational achievement test in reading was 45.00 with a σ of 6.00. At the end of the school year, the mean score on an equivalent form of the same test was 50.00 with a σ of 5.00. The correlation between scores made on the initial and final testing was .60. Has the class made significant progress in reading during the year?

We may tabulate our data as follows:

	Initial Test	Final Test
No. of children:	65	65
Mean score:	45.00(M_1)	50.00(M_2)
Standard deviations:	6.00(σ_1)	5.00(σ_2)
Standard error of the mean:	.74(σ_{M_1})	.62(σ_{M_2})
Difference between means:		5.00
Correlation between initial and final tests.		.60

Substituting in formula (39), we get

$$\begin{aligned}\sigma_D &= \sqrt{(.74)^2 + (.62)^2 - 2 \times .60 \times .74 \times .62} \\ &= .62\end{aligned}$$

* A discussion of the meaning and use of the coefficient of correlation will be found in Chapter IX.

† The correlation between the means of samples drawn from a given population equals the correlation between the test scores, the means of which are being compared. See Kelley, T. L., *Statistical Method*, 1923, p. 178.

and D/σ_D is 5.00/.62 or 8.1. From this result (Table 34) it is clear that the class made significant progress in reading during the school year.

The formula for the PE_D when the means are correlated is

$$PE_D = \sqrt{PE_{M_1}^2 + PE_{M_2}^2 - 2r_{12}PE_{M_1}PE_{M_2}} \quad (40)$$

(probable error of a difference between correlated means)

in which PE_{M_1} and PE_{M_2} are the probable errors of the initial and final test means, and r_{12} is, as before, the correlation between the scores made upon the initial and final tests. The PE_D is interpreted by means of Table 35.

Formulas (39) and (40) are often employed in experiments which make use of the method of "equivalent groups." * The value of the "equivalent groups" method is that it enables us to estimate the effect of some experimentally varied condition ("experimental factor") as compared with the absence of the factor ("control"), or with some other experimental factor. The following problem is typical of many which employ the equivalent group technique.

Problem (4) Two groups (A and B) of seventh grade children are matched, child for child, for age and for score upon Form A of the Otis Group Intelligence Scale (Advanced Examination). Three weeks later, both groups were given Form B of the same test. Before the second test, however, Group A, the "experimental group," was praised for its performance on the first test, and urged to better its score if possible. Group B, the "control group," was simply given the test without comment. Will the incentive (praise) serve to increase significantly the final mean score of Group A over Group B?

* Walker, H. M., *Concerning the Standard Error of a Difference*, Journal of Educational Psychology, 1929, 20, pp. 53-60.

The relevant data may be tabulated as follows:

	Experimental Group A	Control Group B
No. of children in each group:	72	72
Mean scores on Form A, initial test:	80.42	80.51
<i>SD</i> on Form A, initial test:	23.61	23.46
Mean scores on Form B, final test:	88.63(M_1)	83.24(M_2)
<i>SD</i> on Form B, final test:	24.36(σ_1)	21.62(σ_2)
Gain, $M_1 - M_2$.	5.39	
Standard errors of means, final tests:	2.87	2.55

Correlation between final scores (experimental and control groups): .65

The means and σ 's of the control and experimental groups are almost identical, showing how closely the matching has been made. The correlation between the final scores on Form B of the Otis Test is found by pairing off the scores of those children who were matched in terms of initial score.*

The difference (D) in the final mean test performance of the experimental and control groups is $88.63 - 83.24$ or 5.39 . The standard error of this D , σ_D , is found from formula (39) as follows:

$$\begin{aligned}\sigma_D &= \sqrt{(2.87)^2 + (2.55)^2 - 2 \times .65 \times 2.87 \times 2.55} \\ &= 2.29\end{aligned}$$

Hence, D/σ_D is $5.39/2.29$ or 2.35 , and from Table 34, we find that there are 99 chances in 100 that the superiority of the "incentive" group is significant. Had no account been taken of the correlation between the two sets of final scores in the control and experimental groups, i.e., if formula (37) had been used, the σ_D would have been 3.84; and D/σ_D would have been 1.40 instead of 2.35. The chances of a significant difference between the final mean scores by formula (37) are 92 in 100, instead of 99 in 100, the correct result.

* Note that the correlation between the final scores of two matched groups is really equivalent to the correlation between initial and final scores of the same group. The control group furnishes the "initial" scores — the standard of comparison.

The method of "equivalent groups" has been employed in a variety of psychological and educational studies. Well known illustrations are found in experiments designed to evaluate the relative merits of two methods of teaching; to determine the effects of drugs, e g., tobacco or caffeine, upon efficiency; to investigate the transfer effects of special training, and many other factors. When groups are not matched subject for subject in initial performance, but are simply equated as to mean and σ by a random selection of fairly large samples, it is impossible to calculate the correlation between final scores. Formula (37) is then usually employed although it gives results which may be subject to considerable error. If D/σ_D , when formula (37) is used, is found to be significant, we may put confidence in our result since the D/σ_D given by formula (37) is *always smaller* than the ratio given by formula (39) when r is positive and larger than zero. But if the difference when formula (37) is used is *not* significant, we cannot be sure that it might not prove to be reliable if the experiment were repeated under conditions changed so as to permit the calculation of the correlation between the final scores in experimental and control groups. Often groups are matched in some general ability such as intelligence or educational achievement, and their end performances in some different function (some function other than that used in matching) are compared. In such cases various special formulas devised for these particular situations should be employed.*

2. The Reliability of the Difference Between Medians

It is often necessary to find the reliability of the difference between median scores made on the same test by different groups. This may be done readily by rewriting formulas (37) and (38), used in evaluating the difference between two uncorrelated means. Thus

$$\sigma_D = \sigma_{M\bar{a}n_1 - M\bar{a}n_2} = \sqrt{\sigma^2_{M\bar{a}n_1} + \sigma^2_{M\bar{a}n_2}} \quad (41)$$

* Lindquist, E. F., *The Significance of a Difference Between "Matched" Groups*, Journal of Educational Psychology, 1931, 22, pp. 197-204.

Wilks, S. S., *The Standard Error of the Means of "Matched" Samples*, Journal of Educational Psychology, 1931, 22, pp. 205-208.

$$\text{and } PE_D = PE_{Mdn_1 - Mdn_2} = \sqrt{PE_{Mdn_1}^2 + PE_{Mdn_2}^2} \quad (42)$$

*(standard and probable errors of the difference between
medians obtained from different groups)*

in which σ_{Mdn} is found by formula (30) or formula (33).

We may illustrate the use of these formulas with a problem.

Problem (5) The following results were obtained upon the Trabue Language Scale A* from two groups of 12 year old boys and 12 year old girls in Grades III to VIII.

	Boys	Girls
<i>N</i>	801	448
Median	21.40	22.80
<i>Q</i>	4.9	5.3

The difference between the two median scores is 1.4 in favor of the girls. On the assumption that the two groups are unselected, normally distributed, and fairly representative of their populations, is this difference large enough to insure that the true difference is greater than zero and is in favor of the girls? Since the measure of variability given by the problem is the *Q*, we shall calculate the PE_{Mdn} 's as measures of reliability for the two medians. Thus —

$$\text{For girls: } PE_{Mdn} = 1.2533 \times \frac{5.3}{\sqrt{448}} = .31 \quad \text{by (32)}$$

$$\text{For boys: } PE_{Mdn} = 1.2533 \times \frac{4.9}{\sqrt{801}} = .22 \quad \text{by (32)}$$

Substituting in formula (42), we have

$$PE_D = \sqrt{(.31)^2 + (.22)^2} = .38$$

The obtained difference is 1.4 and the PE_D is .38. Therefore, D/PE_D is 3.7, and from Table 35 the chances are 99 in 100 that the obtained difference could not have arisen by chance. The difference of 1.4 is about 90% (3.7/4.0) of what it should be conventionally in order to guarantee complete reliability.

* Trabue, M. R., *Completion Test Language Scale*, Teachers College, Columbia University, Contributions to Education, 1916, 77, p. 15.

However, it is large enough to be strongly indicative of a significant difference in favor of the girls.

3. The Reliability of the Difference Between Standard Deviations

(1) The Standard Error of a Difference When σ 's Are Uncorrelated ($\sigma_{D\sigma}$)

In many studies in psychology and education, the differences in variability which appear between groups are a matter of prime importance. The student of race and sex differences, for example, is often much more interested in whether or not his groups differ significantly in variability than in whether or not they differ in mean score. Likewise, the educational psychologist, who is investigating a new method of teaching, often wants to know whether his new method has produced changes in variability in the final scores of his experimental group which are reliably greater than the changes in variability in the final scores of his control group (taught by the "old" method).

When the difference in the variability of two *different* groups* is studied, so that there is no correlation between the σ 's computed from successive samples, the reliability of an obtained difference may be measured by the formula

$$\sigma_{D\sigma} = \sigma_{\sigma_1 - \sigma_2} = \sqrt{\sigma^2_{\sigma_1} + \sigma^2_{\sigma_2}} \quad (43)$$

(standard error of the difference between uncorrelated σ 's)

where σ_{σ_1} is the standard error of the first σ (σ_1), and σ_{σ_2} is the standard error of the second σ (σ_2) [see formula (34) for σ_{σ}].

We may apply this formula to the problem on page 211. The σ of the Germans' scores on the "combined scale" is 2.43; of the Danes' scores on the same test, 2.23. Is the difference in variability between these two groups too great to be explained by chance? Calling the σ of the Germans' scores σ_1 and the σ of the Danes' scores σ_2 , we have

* Formula (43) may be used, also, when tests given to the same group are uncorrelated.

$$\sigma_{\sigma_1} = \frac{2.43}{\sqrt{616}} = .098 \quad \text{by formula (34)}$$

$$\sigma_{\sigma_2} = \frac{2.23}{\sqrt{650}} = .087 \quad \text{by formula (34)}$$

If now we substitute these values in formula (43), the

$$\begin{aligned}\sigma_{D_\sigma} &= \sqrt{(.098)^2 + (.087)^2} \\ &= .131\end{aligned}$$

The obtained difference in the σ 's is $2.43 - 2.23$ or $.20$. Dividing $.20$ (D) by $.131$ (σ_{D_σ}), we obtain 1.53 ; and from Table 34, find that there are 93 chances in 100 that the true difference in variability as between the scores of Germans and Danes is greater than zero. While the obtained difference in variability between these two groups is not quite significant, we do know that 9 times in 10 the σ of the Germans' scores will be larger than the σ of the Danes' scores.

(2) The Standard Error of a Difference When σ 's Are Correlated

We have already found that when the difference between the mean scores of the same group or of matched groups is investigated, we must take account of the correlation between the means of the two groups compared (p. 217). In like manner, when the same group or when matched groups are studied for differences in variability, we must take account of the correlation between the σ 's of the groups compared. The formula for testing the significance of an obtained difference in variability when the σ 's are correlated is

$$\sigma_{D_\sigma} = \sqrt{\sigma^2_{\sigma_1} + \sigma^2_{\sigma_2} - 2r^2_{12}\sigma_{\sigma_1}\sigma_{\sigma_2}} \quad (44)$$

(standard error of the difference between correlated σ 's)

where σ_{σ_1} and σ_{σ_2} are the standard errors of the two σ 's and r^2_{12} * is the square of the coefficient of correlation between the

* The correlation between the σ 's of samples drawn from a given population equals the square of the coefficient of correlation between the test scores, the σ 's of which are being compared. See Kelley, T. L., *Statistical Method*, 1923, p. 178.

scores in initial and final tests of the same group, or between the final scores of two matched groups.

Formula (44) may be applied to the problems on pages 218 and 220. In the first problem (p. 218), the σ of the 65 sixth grade children is 6.0 on the initial test and 5.0 on the final test. Is there a significant drop in variability in reading after a year's schooling? If we let $\sigma_1 = 6.0$, and $\sigma_2 = 5.0$, we have

$$\sigma_{\sigma_1} = \frac{6.0}{\sqrt{130}} = .53 \quad \text{by formula (34)}$$

$$\sigma_{\sigma_2} = \frac{5.0}{\sqrt{130}} = .44 \quad \text{by formula (34)}$$

The coefficient of correlation between initial and final scores is .60, so that $r^2 = .36$. Substituting for r^2 and our calculated values in formula (44) we have

$$\begin{aligned} \sigma_{D\sigma} &= \sqrt{(.53)^2 + (.44)^2 - 2 \times .36 \times .44 \times .53} \\ &= .55 \end{aligned}$$

The difference of 1.0 (D) divided by .55 ($\sigma_{D\sigma}$) equals 1.8; and from Table 34 the chances are 96 in 100 that there has been a significant decrease in variability in the reading scores in this group. The change in variability is not completely reliable (p. 213), but the evidence for a decrease is strong, and, if substantiated by other experiments, might be taken as conclusive.

In the second problem, which involves matched groups (p. 220), the σ of the experimental group on the final test is 24.36; and the σ of the control group on the final test 21.62. The number of children in both groups is 72. Did the incentive (praise) produce significantly greater variability in the experimental group as compared with the control? Putting $\sigma_1 = 24.36$, and $\sigma_2 = 21.62$, we have

$$\sigma_{\sigma_1} = \frac{24.36}{\sqrt{144}} = 2.03 \quad \text{by (34)}$$

$$\sigma_{\sigma_2} = \frac{21.62}{\sqrt{144}} = 1.80 \quad \text{by (34)}$$

The coefficient of correlation between the final test scores of the experimental and control groups is .65, and r^2_{12} is .42. Substituting for r^2 and our calculated values in formula (44), we have

$$\begin{aligned}\sigma_{D\sigma} &= \sqrt{(2.03)^2 + (1.80)^2 - 2 \times .42 \times 2.03 \times 1.80} \\ &= 2.07\end{aligned}$$

If we divide 2.74 (24.36 - 21.62) by 2.07, our $D/\sigma_{D\sigma}$ equals 1.32; and from Table 34 there are 90 chances in 100 that the experimental group is significantly more variable than the control. The evidence for greater variability in the experimental group is not conclusive, but it is strong enough to guarantee that 9 times in 10 we should find greater variability in our experimental groups.

V. THE RELIABILITY OF CERTAIN OTHER MEASURES

In this section, we shall consider the standard errors of certain calculated measures which occur fairly often in ordinary statistical work. The *PE* of r , the coefficient of correlation, will be discussed in Chapter IX, page 281. For the standard errors and *PE*'s of many other important measures the student should go to the more advanced references in the bibliography. The *Handbook of Statistical Nomographs, Tables, and Formulas*, by Dunlap and Kurtz, contains many formulas which are often necessary in research (p. 465).

1. The Standard Error of a Percentage and the Standard Error of the Difference Between Two Percentages

It is often possible to find the percentage of a given group which exhibits a certain attribute or possesses certain interests or attitudes, or other fairly general characteristics, when it is difficult if not impossible to measure these attributes directly. Given the percentage occurrence of an attribute, the question often arises of how much confidence we can put in our figure. How reliable an index is it of the incidence of the phenomenon

in which we are interested? The standard error of a percentage is given by the formula

$$\sigma_p = \sqrt{\frac{pq}{N}} \quad (45)$$

(standard error of a percentage)

in which p = the percentage of times the given event occurs; $q = 1 - p$, and N = the number of cases.

We may illustrate this formula with a problem:

Problem (1) In a study of deception or cheating, a group of 613 elementary school children were classified as to the occupations of their fathers. It was found that 348 children had fathers who were professional men, business men, merchants, etc. Of these 348 children of "good" social status, 144 or 41.4% were found to have cheated on various tests given in school. Assuming our sample to be representative of children from the given social level, how much confidence may be placed in the stability of this percent? How much fluctuation in percent cheating might be expected if we investigated a number of groups of children whose fathers fell into the same occupational classification? *

Applying formula (45), we get

$$\sigma_p = \sqrt{\frac{.414 \times .586}{348}} = .027 \text{ or approximately } 3\%$$

This standard error is interpreted as is σ_M . If our group is "typical," i.e., fulfills the conditions of random sampling, the chances are 68 in 100 that the "true" percentage † of children from "good" social levels who cheat on these tests will fall within the limits $.414 \pm .027$ or between 39% and 44%. We may be reasonably sure that the percentage cheating will

* Hartshorne, H., and May, M. A., *Studies in Deceit*, 1928, Book 2, p. 161.

† The true percentage may be thought of as the average of all of the percents obtained when successive samples are drawn from the parent population, and the percent cheating in each sample determined.

not be greater than 50% nor less than 33% (will lie within $\pm 3\sigma_p$).*

We often want to know whether there is a significant difference between the percentages of two groups who exhibit a certain form of behavior or show a certain attribute. When our two groups constitute samplings from different populations, we may compute the reliability of the difference between the percentages in the two groups exhibiting the given form of behavior by the formula:

$$\sigma_{D_p} = \sigma_{p_1 - p_2} = \sqrt{\sigma_{p_1}^2 + \sigma_{p_2}^2} \quad (46)$$

$$\text{or} \quad \sigma_{D_p} = \sqrt{\frac{p_1 q_1}{N_1} + \frac{p_2 q_2}{N_2}}$$

(standard error of the difference between two uncorrelated percentages)

We may illustrate the use of this formula by reference to the problem of cheating given above. It was there stated that 41.4% of the 348 children classified as of "good" social status cheated on the tests given. In the same study, 50.2% of 265 children whose fathers were classified as skilled and unskilled laborers, i.e., were of "poor" social status, cheated on the same tests of deception. Is there a "real" difference in "deceptive behavior" between these two groups? The σ_p for the percentage .502 in the second group is

$$\sigma_p = \sqrt{\frac{502 \times .498}{265}} = .031 \text{ or approximately } 3\%$$

Calling .027 σ_{p_1} , and .031 σ_{p_2} , and substituting in formula (46), we have

$$\sigma_{p_1 - p_2} = \sqrt{(.027)^2 + (.031)^2} = .041$$

The difference between the percentage cheating in the two groups is .502 - .414 or .088. Dividing .088 by .041, we ob-

* For a convenient table giving the range within which the probability is .99 that a given percent falls, see Hart, H., *The Reliability of a Percentage*, Journal of the American Statistical Association, 1926, 21, pp. 40-46.

tain a D/σ_{D_s} of 2.15; and from Table 34, we find that the chances are 98 in 100 that, on the average, children from poorer social levels will cheat to a greater extent than will children of better social status.

2. The Standard Errors of Measures of Skewness and of Kurtosis

(1) Skewness

In Chapter V, page 116, a formula for estimating the skewness of a frequency distribution in terms of its median and certain percentiles was given as follows:

$$Sk = \frac{(P_{90} + P_{10})}{2} - P_{50} \quad \text{formula (19)}$$

According to this formula, the skewness of the 50 Army Alpha scores, whose distribution is given in Table 1, page 5, is -2.50 . The significance of this measure of skewness may be calculated by means of the formula

$$\sigma_{sk} = \frac{.5185 D}{\sqrt{N}} \quad (47)$$

(standard error of the measure of skewness given in formula (19) *)

in which $D = (P_{90} - P_{10})$.

In the frequency distribution of 50 Army Alpha scores, P_{90} is 187, P_{10} is 152, and $D = 35$. From formula (47), therefore,

$$\sigma_{sk} = \frac{.5185 \times 35}{\sqrt{50}} = 2.57$$

and dividing -2.50 (Sk) by 2.57 (σ_{sk}), we get $-.97$ as the Sk/σ_{sk} value. From Table 34 it appears that there are 83 chances in 100 that the obtained Sk is significant. The departure of this frequency distribution from normality ($Sk = .00$), therefore, while fairly large, is not marked.

* Kelley, T. L., *Statistical Method*, 1923, p. 77. The formula, as given in this reference, is in error; see Dunlap, J. W., and Kurtz, A. K., *Handbook of Statistical Nomographs, Tables and Formulas*, 1932, p. 112.

The skewness of the distribution of 200 Cancellation scores (p. 19) is, by formula (19), .03; $P_{90} = 128.5$, $P_{10} = 110.4$, and $D = 18.1$. The standard error of Sk is

$$\sigma_{sk} = \frac{.5185 \times 18.1}{\sqrt{200}} = .664$$

Dividing .03 (Sk) by .664 (σ_{sk}), we get .05, and from Table 34 find that there are only 52 chances in 100 that the skewness is significant. This distribution, therefore, is almost perfectly symmetrical (Fig. 5, p. 71, verifies this result).

(2) Kurtosis

On page 118 the following formula for measuring the kurtosis of a distribution in terms of Q and certain percentiles was given:

$$Ku = \frac{Q}{(P_{90} - P_{10})} \quad \text{formula (20)}$$

The kurtosis of the frequency distribution of 50 Army Alpha scores by formula (20) is given on page 118 as .237. This value deviates .026 from the Ku of the normal probability distribution which is .26315. The direction of the deviation indicates that the distribution is slightly leptokurtic.

We may estimate the significance of our deviation of .026 from "normal" kurtosis by calculating σ_{Ku} , using the following formula

$$\sigma_{Ku} = \frac{.27779}{\sqrt{N}} \quad (48)$$

(*standard error of the measure of Ku given by formula (20)*)

in which N is the size of the sample.

For the 50 Army Alpha scores, $\sigma_{Ku} = \frac{.27779}{\sqrt{50}} = .0393$, and

$Ku/\sigma_{Ku} = .026/.0393$ or .66. There are, therefore, 74 chances in 100 that the deviation of this frequency distribution from the normal form (its "peakedness") is significant.

The kurtosis of the 200 Cancellation scores (p. 19) is by formula (20) .223, a value which deviates .040 from .26315, the Ku of the normal distribution. To estimate the significance of

this deviation from normality, we first calculate σ_{Ku} which equals .0196. Ku/σ_{Ku} equals .040/.0196 or 2.04, and from Table 34 we find that there are 98 chances in 100 that the deviation of this distribution from the normal form is significant. The rather narrow dispersion of this distribution ($Q = 4.04$) and its relatively large size (200 cases) probably account for the fact that it exhibits a strong tendency to be more "peaked" than the normal distribution. The distribution, to be sure, is not significantly peaked, judged by conventional standards, as its Ku still lies within the range of $\pm 3\sigma_{Ku}$ from the Ku of the normal distribution. Since the distribution is not skewed, it may, therefore, be profitably treated as normal despite its decided tendency toward leptokurtosis.

VI. VARIOUS PROBLEMS WHICH INVOLVE MEASURES OF RELIABILITY

This section illustrates a number of problems which require for their solution a knowledge of the reliability formulas given in this chapter, and the ability to use and interpret the probability tables (Tables 15 and 16). For later reference, each group of examples is preceded by a general statement of the essential issues involved.

1. To Find the Probability that the True Measure is Greater or Less than Some Designated Point on the Scale, or that it Falls Within Certain Given Limits.

Problem (1) Given a mean of 30.20, a σ of 6.0, and an $N = 100$ On the assumption that this sample is normally distributed and representative of the population from which it is drawn, (a) what is the reliability of the obtained mean? (b) What are the chances that the true mean is less than 29.00? (c) 31.50 or greater? (d) that the true mean lies between 28.00 and 31.00? (e) that the true σ lies between 5.0 and 7.0?

(a) From formula (28) we find that σ_M is .60. Hence, from Table 15 the chances are 68 in 100 that this mean does not

diverge from the true mean by more than $\pm .60$ — or, that the true mean falls between the limits 29.60 and 30.80. Furthermore, the chances are 997 in 1000 that 30.20 does not diverge from the true mean by more than $\pm .60 \times 3$ or ± 1.80 ; or that the true mean falls within the limits 28.40 and 32.00.

These results will be clearer, perhaps, if represented graphically as in Figure 39. This normal distribution pictures the distribution of *means* which we should expect to get theoretically

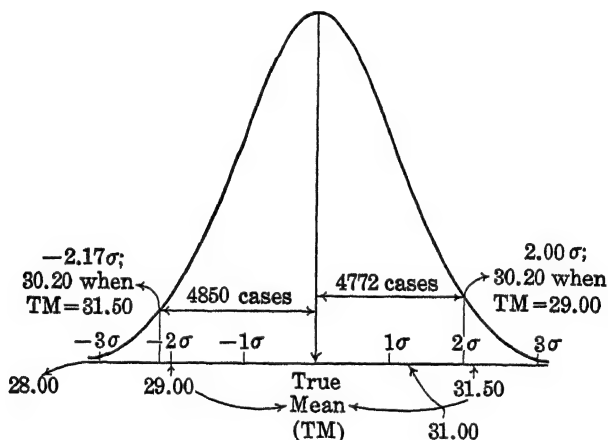


FIG. 39.

from a large number of samples drawn at random from the population in the same way as was the present sample whose mean is 30.20. The mean of this hypothetical distribution-of-means is the "true mean" (*TM*) whose value we are trying to estimate; and the best estimate of the σ_M of this distribution is .60, the standard error of the given obtained mean

(b) What are the chances that the true mean is less than 29.00? A mean of 29.00 is 1.20 points or 2.0σ ($1.20/.60 = 2.0$) below our obtained mean of 30.20. Now what we want to know is the probability that we should have obtained a mean of 30.20, if 29.00 were *actually* the true mean (*TM*), i.e., occupied the middle of the baseline. From Figure 39 it is clear that if

29.00 fell at the true mean, 30.20 would be 2.0σ *above* the true mean. From Table 15, we find that 4772 cases in 10,000 fall *between* the mean and 2σ in a normal distribution; and that 5000 - 4772 or 228 cases lie *above* this point. Hence, there are only 228 chances in 10,000 or about 2 in 100 that we should have obtained the mean we did (namely, 30.20) if the true mean were less than 29.00. Or, the chances are 2 in 100 that the true mean is less than 29.00.

(c) What are the chances that the true mean is as large as or larger than 31.50? A mean of 31.50 is 1.30 points or 2.17σ *above* our obtained mean of 30.20. This time we want to know the probability that we should have obtained a mean of 30.20, if 31.50 were actually the true mean. From Figure 39 it is clear that if 31.50 fell at the true mean (*TM* or middle of the baseline) 30.20 would be 2.17σ *below* this point. From Table 15, we find that 4850 cases in 10,000 fall *between* the mean and -2.17σ in a normal distribution; and that 5000 - 4850 or 150 cases fall *below* -2.17σ . Hence, there are only 150 chances in 10,000 that we should have obtained our mean of 30.20, if 31.50 or a larger value were, in fact, the true mean. Or, the chances are only 15 in 1000 that the true mean is 31.50 or more.

(d) What are the chances that the true mean lies between 28.00 and 31.00? A mean of 28.00 is 2.20 or 3.67σ *below* our obtained mean of 30.20; and a mean of 31.00 is .80 or 1.33σ *above* the obtained mean of 30.20. What is the probability that we should have obtained a mean of 30.20, if the true mean were somewhere between 28.00 and 31.00? Again from Table 15, we find that 4999 cases in 10,000 lie between the mean and 3.67σ in a normal distribution; and that 4082 cases in 10,000 lie between the mean and 1.33σ . If 28.00 moved up to the middle of the distribution (became the true mean), 30.20, our obtained mean, would be 3.67σ *above* the true mean; and if 31.00 were the true mean, 30.20 would be 1.33σ *below* the true mean. Hence, if the true mean falls *anywhere* between these two points, there are 4999 + 4082 or 9081 chances in 10,000 that we should have obtained our mean of 30.20. Stated

differently, there are 91 chances in 100 that 30.20 would have been obtained, *if* the true mean were, in fact, between 28.00 and 31.00; hence this is the probability (91 in 100) that the true mean lies between 28.00 and 31.00.

(e) What are the chances that the true σ lies between 5.0 and 7.0? This is essentially the same problem as in (d) above. From formula (34) we find σ_σ to be .42. A σ of 5.0 is 1.0 or 2.38σ below the obtained σ of 6.0; and a σ of 7.0 is 1.0 or 2.38σ above the obtained σ of 6.0. What is the probability that we should have obtained a σ of 6.0, if the true σ lay somewhere between 5.0 and 7.0? From Table 15, we find that between the mean of a normal distribution and $+2.38\sigma$ or -2.38σ are 4913 cases. Now, if 5.0 were the true σ , our obtained σ of 6.0 would be 2.38σ above the true σ ; and if 7.0 were the true σ , our obtained σ of 6.0 would be 2.38σ below the true σ . There are 4913×2 or 9826 chances in 10,000, therefore, or 98 in 100, that if the true σ fell *anywhere* between these two values, we should have obtained our σ of 6.0. Or the chances are 98 in 100 that the true σ lies between 5.0 and 7.0.

Problem (2) Given a median of 26.4, and a PE_{Mdn} of 1.5. What are the chances that the true median of the population, of which the given group is a random sampling, is (a) as large as 30.0? (b) as small as 24.0?

As in problem (1) above, the hypothetical distribution of medians from successive samples may be represented by a normal curve with the true median at the center of the distribution and PE_{Mdn} equal to 1.5. The probable error of the obtained median is the best estimate we possess of the probable error of this distribution.

(a) What are the chances that the true median is 30.0 or more? A median of 30.0 is 3.6 score points or $2.4PE$ above our obtained median of 26.4. What we want to know is the probability that we should have obtained a median of 26.4, if 30.0 or a larger value were actually the true median. From Figure 40, it is clear that if 30.0 were actually the true median

(fell at the middle of the distribution), 26.4 would be $2.4PE$ below the true median. From Table 16, we find that 4473 cases in 10,000 in a normal distribution lie between the mean and $2.4PE$; and that 5000 - 4473 or 527 cases lie below $2.4PE$. Hence, there are only 527 chances in 10,000, or about 5 in 100, that we should have obtained a median of 26.4, if 30.0 or a larger value were actually the true median. Or the chances are only 5 in 100 that the true median is 30.0 or more.

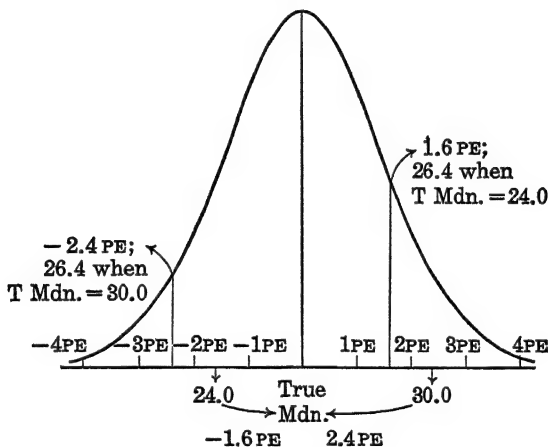


FIG. 40.

(b) What are the chances that the true median is as small as 24.0? A median of 24.0 is 2.4 score points or 1.6*PE* below our obtained median of 26.4. This time we ask ourselves what the probability is that we should have obtained a median of 26.4, if the true median were 24.0 or less. From Figure 40, it may be seen that if 24.0 were the true median (moved up to the center of the distribution), 26.4, our obtained median, would then be 1.6*PE* above the true median value. From Table 16, we find that 3597 cases in 10,000 in a normal distribution fall between the median and 1.6*PE*; and that 1403 cases lie above this point. Hence, there are 1403 chances in 10,000, or 14 in 100, that we should have obtained our median of 26.4, if the

true median were actually as small as 24.0. Or the chances are only 14 in 100 that the true median is as small as 24.0 (or smaller).

2. To Find the Probability that the Divergence of an Obtained Measure from its True Measure Will Lie Within Given Limits

Problem (3) Given a mean of 152.7 and a σ_M of 4.5. Find the probability that the given mean does *not* diverge from the true mean by more than (a) ± 1.0 , (b) ± 3.0 , (c) ± 5.0 , and (d) ± 10.0 .

(a) This is essentially the same problem as (1) above, expressed in a slightly different way. In order to find the probability that the obtained mean diverges from the true mean by as much as ± 1.0 , we must find the probability that a mean of 152.7 would have arisen, if the true mean lay somewhere within the limits 151.7 and 153.7. A mean of 151.7 is 1.0 score unit or $.22\sigma$ ($1/4.5 = .22$) below the obtained mean of 152.7; and a mean of 153.7 is 1.0 score unit or $.22\sigma$ above the obtained mean of 152.7. From Table 15, we find that 871 cases in 10,000 in a normal distribution fall between the mean and $+.22\sigma$ and 871 cases in 10,000 fall between the mean and $-.22\sigma$. Accordingly, 871×2 or 1742 cases fall within the interval $+.22\sigma$ and $-.22\sigma$. Now if 151.7 moved up to the middle of the distribution (became the true mean), our obtained mean would be $.22\sigma$ above the true value; and if 153.7 moved down to the middle of the distribution (became the true mean) our obtained mean would be $-.22\sigma$ from the true value. Hence, if the true mean falls *anywhere* between 151.7 and 153.7 there are 871×2 or 1742 chances in 10,000 that we should have obtained the given mean, namely, 152.7. Or, the chances are 1742 in 10,000 that the obtained mean does not differ from the true mean by more than ± 1.0 .

(b) A mean 3 score units above or below the obtained mean lies $\pm 3/4.5$ or $\pm .67\sigma$ from the obtained mean. In a normal distribution 2486×2 or 4972 cases in 10,000 fall within the interval $\pm .67\sigma$ from the mean. Hence, the chances are 50

in 100 that we should have obtained the mean we did, namely, 152.7, if the true mean falls between 149.7 and 155.7; or that the obtained mean does not diverge from the true mean by more than ± 3.0 .

(c) A mean 5 score units above or below the obtained mean lies $\pm 5/4.5$ or $\pm 1.1\sigma$ from this value. In a normal distribution 3643×2 or 7286 cases fall within the interval $\pm 1.1\sigma$ from the mean. Hence, the chances are 73 in 100 that we should

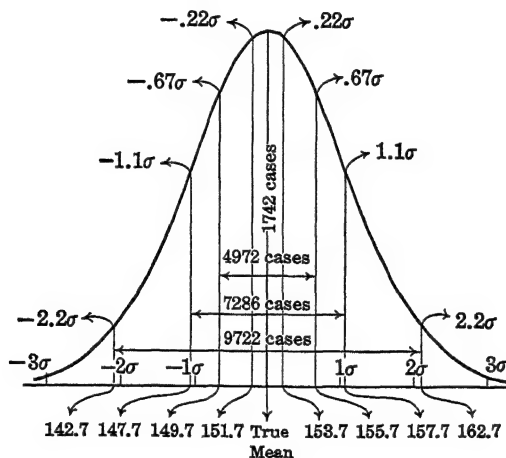


FIG. 41

have obtained a mean of 152.7, if the true mean actually lies between 147.7 and 157.7; or that the obtained mean does not diverge from the true mean by more than ± 5.0 .

(d) A mean 10 score units above or below the obtained mean lies $\pm 10/4.5$ or $\pm 2.2\sigma$ from this value. In a normal distribution (Table 15) 4861×2 or 9722 cases in 10,000 fall within the interval $\pm 2.2\sigma$ from the mean. Hence, the chances are 97 in 100 that we should have obtained a mean of 152.7, if the true mean lies somewhere between 142.7 and 162.7; or that the obtained mean does not diverge from the true by more than ± 10.0 .

Problem (4) In a group of 64 children, each of whom has one parent a church member, and one not a church member, 45% of the children are themselves church members. On the assumption that our sample is a fair selection from a larger population, what is the probability that the obtained percentage diverges from the true percentage (a) by 2% or more? (b) by 10% or more?

(a) Applying formula (45), we obtain a σ_p of .06. Now, in order to find the probability that our obtained percentage of .45 diverges from the true percentage by 2% or more, we must first find the probability that a percentage of .45 would have been obtained, if the true percentage lies between .43 and .47. A percentage of .43 is $-.02$ or $-.33\sigma$ (.02/.06) from the obtained percentage; and a percentage of .47 is $+.02$ or $+.33\sigma$ from the obtained percentage. From Table 15, we find that 1293×2 or 2586 cases in 10,000 in a normal distribution fall within the interval $\pm .33\sigma$ from the mean. Hence, the chances are 26 in 100 that we should have obtained 45%, if the true percentage lay between 43% and 47%; and the chances are, of course, 74 in 100, that we should have obtained 45% if the true percentage were .43 or less or if it were .47 or more. Thus, we may say that the chances are 74 in 100 that the obtained percentage *does* diverge from the true percentage by 2% or more.

(b) In order to find the probability that our obtained percentage of .45 diverges from the true percentage by 10% or more, we shall again find first the probability that a percentage of .45 would arise, if the true percentage lay between .35 and .55. A percentage of .35 is $-.10$ or -1.67σ from the obtained percentage of .45 ($-.10/.06 = -1.67$); and a percentage of .55 is $.10$ or 1.67σ above the obtained percentage. From Table 15, we find that 4525×2 or 9050 cases fall within the interval $\pm 1.67\sigma$ from the mean in a normal distribution. Hence, the chances are 91 in 100 that we should have obtained a percentage of .45, if the true percentage lay between 35% and 55%. Or, answering the question above, the chances are 9 in 100 that

we should have obtained a percentage of .45, if the true percentage were .35 or less, or if it were .55 or more. Hence, there are only 9 chances in 100 that the given obtained percentage of .45 diverges 10% or more from the true percentage.

3. To Find the Probability that the True Difference Between the Measures of Two Groups is Greater or is Less than a Given Amount

Problem (5) The difference between two obtained means is 3.0, and the standard error of this difference (σ_D) is 1.5.

(a) What are the chances that the true difference between the means of our populations is greater than zero? (b) Is greater than 1.0? (c) Greater than 3.0?

(a) The situation here is represented graphically in Figure 42. This normal curve shows the distribution of differences which we should expect theoretically if a large number of samples, drawn at random from our populations, were compared two at a time. The mean of this hypothetical distribution-of-differences is set at zero, and the σ_D is taken to be 1.5 (our best estimate of its value). Our problem now is to find the probability that we should have obtained a difference of 3.0, if the true difference were actually zero. Our obtained difference of 3.0 is 3.0 score units or 2.0σ ($3/1.5 = 2.0$) from the hypothetical true zero. From Table 15, we find that 4772 cases in a normal distribution fall between the mean and 2.0σ ; and that 228 cases fall above this point. Hence, on the assumption that the true difference is zero, there are only 228 chances in 10,000, or 2 in 100, that we should have obtained a difference as large as (or larger than) 3.0. Conversely, there are 98 chances in 100 that we should *not* have obtained a difference of 3.0, if the true difference were zero. Hence, we can say that the chances are 98 in 100 that the true difference is somewhat greater than zero; and 2 in 100 that the true difference is zero or negative. The odds against the hypothesis that the true difference is zero, therefore, are 49 to 1.

(b) What are the chances that the true difference is greater than 1.0? This question may be restated as follows: what is

the probability that we should get a difference of 3.0, if the true difference were greater than 1.0? Our obtained difference of 3.0 is 2.0 score units, or 1.3σ ($2.0/1.5$), above the true zero, now assumed to be 1.0 (see Fig. 42). From Table 15, we know that in a normal distribution there are 4032 cases in 10,000 between the mean and 1.3σ ; and that 968 cases lie above this point. On the assumption that the true difference is 1.0, or less, there are 968 chances in 10,000, or about 1 in

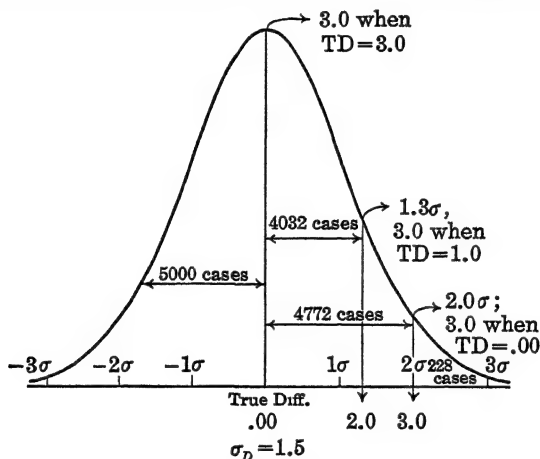


FIG. 42.

10, that we should have obtained a difference as large as 3.0. And there are 9032 ($5000 + 4032$) chances in 10,000, or 90 in 100, that we should have obtained a difference of 3.0, if the true difference were more than 1.0. Hence, there are 90 chances in 100 that the true difference is greater than 1.0.

(c) What are the chances that the true difference is greater than 3.0? If the true difference were 3.0, our obtained difference of 3.0 would coincide with it exactly. Thus, 3.0 would fall at the center of the baseline, and the chances would be 50 in 100 (the odds would be even) that the true difference is 3.0 or more. The chances likewise would be 50 in 100 that the true difference is less than 3.0.

Problem (6) On a final test, the σ of an experimental group (σ_1) is 10.64, and the σ of the control group (σ_2) is 8.53. The coefficient of correlation between the final scores of the two matched groups is .42. N is 125 in both groups. What are the chances that the true difference between the σ 's of the experimental and control groups is (a) greater than zero? (b) less than zero? (c) greater than 2.0?

(a) By formula (44) $\sigma_{D\sigma} = .78$, and D (10.64 - 8.53) is 2.11. Our problem is to find the probability that a difference of 2.11 would arise, if the true difference were actually zero. Our difference of 2.11 is $2.11/.78$ or 2.7σ from the true difference of zero. From Table 15, we find that 4965 cases in 10,000 fall between the mean and 2.7σ in a normal distribution; and that 35 cases in 10,000 fall above this point. Hence, there are only 35 chances in 10,000 that a difference of 2.11 would have been found, if the true difference were zero; and 9965 (5000 + 4965) chances in 10,000 that since 2.11 was found the true difference is greater than zero.

(b) This question is answered by (a) above. Since there are 9965 chances in 10,000 that the true difference is greater than zero, there can be only 35 chances in 10,000, or about 4 in 1000, that the true difference is zero or less than zero.

(c) What are the chances that the true difference is more than 2.0? Our obtained difference of 2.11 is .11 score units or $.14\sigma$ ($.11/.78$) above the "true" difference of 2.0. Between the mean and $.14\sigma$ in a normal distribution are 557 cases in 10,000 (Table 15); and above $.14\sigma$ there are 4443 cases in 10,000. Hence, there are 4443 chances in 10,000 that a difference of 2.11 would be obtained if the true difference were 2.0 or less; and 5557 (5000 + 557) chances in 10,000 that 2.11 would be obtained if the true difference were more than 2.0. We may say, then, that the odds are about 56 to 44 that the true difference is somewhat greater than 2.0.

VII. SAMPLING AND RELIABILITY

All of the reliability formulas given in this chapter depend upon N , the number of cases in the sample; and most of them involve, also, some measure of variability (usually σ) calculated from the obtained data. It is unfortunate, perhaps, that given these *parameters*,* there is nothing in the *statement* of a reliability formula itself which might deter the uncritical worker from applying the formula to any set of test scores. The general and indiscriminate calculation of standard errors and probable errors, however, will surely lead to erroneous conclusions and false interpretations. For this reason, it is exceedingly important that the worker in psychology or in education have clearly in mind (1) the conditions under which the reliability formulas are — and are not — applicable; and that he know (2) what his formula may reasonably be expected to do. Some of the limitations to reliability formulas have been pointed out in this chapter and elsewhere in preceding chapters; these statements will now be amplified and certain cautions to be observed in the use of the formulas indicated.

1. Reliability Formulas Assume an Adequate Sample

An adequate sample is one which is truly representative of the population from which it has been drawn. A representative sample should be “randomly selected;” that is, chosen without bias as to able, poor, and mediocre individuals. It may seem paradoxical, but one must be careful to select his sample “randomly;” for example, a random sample of 10 year old boys should not all be chosen from a very poor neighborhood, nor from an exclusive private school, nor from any larger group in which special selection is known to play an important rôle. Mental traits, which have been carefully measured in large groups of individuals, have usually proved to be normally or approximately normally distributed. We may reasonably as-

* The mean, σ , Q , etc., are called *parameters* of the frequency distribution.

sume, therefore, that most of the attributes in which we are interested will follow the normal curve in the general population. Hence, it follows that the samples with which we work should also be normally distributed. When the distribution of ability in the population is "normal," the range covered by samples of different sizes (drawn from the given population) will be approximately as follows:

$N = 10$	Range $\pm 2.0\sigma$
$N = 50$	" $\pm 2.5\sigma$
$N = 200$	" $\pm 3.0\sigma$
$N = 1000$	" $\pm 3.5\sigma$

A range of $\pm 3.5\sigma$ from the mean includes, in a normal distribution, 9995 cases in 10,000 (see Table 15); and the same range includes 99.95% (i.e., 100%) of the cases in a sample of 100. In a sample of 10,000, 5 cases fall outside of the range indicated; in a sample of 100 *all* of the cases lie within the given range. The more extreme the deviation, the less the probability of its occurrence; hence, in small samples wide deviations from the mean cannot appear if the sample is truly representative of a normally distributed population. When working with small samples, deviations far removed from the mean should be discarded much as a laboratory worker throws out measures of reaction time which are obviously premature or delayed.

One of the simplest tests of the adequacy — the "representativeness" — of a sample consists in drawing from the population several other groups of approximately the same size as the given sample. If the measures of central tendency and variability calculated from all of these groups are of nearly the same size, we may be reasonably assured that our sample is representative. If the correspondence is not close, we must continue to add cases until the successive samples give means and σ 's which are closely similar. More information may be secured by this process, with respect to the reliability of our measures of central tendency and variability, than can be obtained from a blanket use of reliability formulas.

2. Reliability Formulas Assume a "Sufficiently Large" Sample

The value of a standard or probable error is conditioned, in part, upon our having a "sufficiently large" sample. If N is less than 25, there is ordinarily little justification in computing reliability measures, unless one knows more about the character of his population than can be told from the sample alone. As we have already seen (p. 201), standard and probable errors vary inversely as the size of the sample; hence, the larger the sample, in general the smaller the standard error. A fairly simple and practical method of deciding whether a sample is "sufficiently" large is to continue adding cases, drawn at random, until the addition of extra cases fails to produce an appreciable change in the mean, median, or σ . When this point is reached, the sample is probably large enough to be taken as representative of its population. But the corollary must be recognized that mere numbers are not in themselves a guarantee of a representative sample.

3. Reliability Formulas Measure Fluctuations Due to Sampling, and to Errors of Measurement *

Standard and probable errors of obtained means, σ 's, etc., measure both (1) sampling errors and (2) errors of measurement, i.e., variable errors in the test scores themselves. An illustration will make the notion of "sampling errors" clearer. We found (p. 202) that the mean height of 8585 adult British males was 67.46 inches with a standard error of .028 inch. This was interpreted to mean that the chances are 997 in 1000 that the true mean height of British males lies between 67.38 and 67.54 inches. Now by "true mean height" we mean the hypothetical average height of *all* British males from whom our group of 8585 is an attempted random sampling. If our group were perfectly representative, its mean would equal the true mean exactly. Except by chance, however, neither this sample nor another similarly selected and approximately of the same

* Errors of measurement ("chance errors") will be discussed in Chapter XI, page 314.

size will represent the entire population perfectly. Furthermore, it is extremely unlikely that the means calculated from successive samples will equal each other. Nevertheless, if the samples are random, and sufficiently large, and if there is no considerable constant error present, these calculated means will tend to vary around the true mean of the population within a comparatively small range. This range is given by the standard and probable error formulas. In interpreting standard and probable errors, we make the assumption that the *parameters* (e.g., means and σ 's) calculated from successive samples are distributed in accordance with the normal probability curve. Uncertainty as to the exact value of a calculated measure arises from the fact that we must, necessarily, work with samples instead of with the whole population. This introduces variations from sample to sample which variations are the so-called "errors of sampling." Such "errors" are to be thought of, however, not as mistakes, failures and the like, but as fluctuations which arise from the fact that no two samples are ever *exactly* alike.

If the standard error of the mean is large, it does not imply necessarily that the obtained mean is affected by large sampling errors. Much of the standard error may be owing to errors of measurement which have not been eliminated.* On the other hand, when errors of measurement and constant errors are known to be negligible, a small standard error indicates clearly that the reliability of our calculated measure is high insofar as fluctuations due to sampling are concerned. In other words, when the standard error is small with reference to the mean, the calculated mean, or σ , say, is a good estimate of the population mean or σ (i.e., true mean or σ).

4. Reliability Formulas do not Measure the Effects of Constant Errors nor the Failure to Get an Adequate Sample

Errors which arise from inadequate sampling are neither detected nor measured directly by reliability formulas. For ex-

* For methods of eliminating "errors of measurement," see page 331.

ample, the standard error of the mean Army Alpha score made by 500 male college students between the ages of 18 and 25 will not tell us that this mean is not representative of the male population within this age-range. College students, however, obviously constitute a highly selected group; and in consequence other samples of 500 drawn at random from the male population between the ages of 18 and 25 will return very different means and σ 's from that of the college group. These fluctuations in mean score and in σ cannot be attributed to sampling errors, since our sample is not representative of the male population within the given age limits. Of course, if our population were restricted to college men, our sample might be entirely adequate.

Reliability formulas are affected by, but do not reveal, constant errors. Constant errors may arise from many sources: familiarity with the test material, fatigue, faulty technique in giving and scoring tests (over- or under-timing are examples), in fact, from a bias of almost any sort. Standard errors calculated for measures subject to such influences, when not definitely misleading, are at best of doubtful value. The careful study of successive samples, retests whenever practicable, care in controlling conditions, and the use of objective checks will eliminate many troublesome and prolific sources of constant error. One should always remember that even the most refined statistical treatment cannot make bad data yield valid results.

SUMMARY:

1. Given a large and representative sample, standard and probable errors of means, σ 's, percents, etc., measure adequately the stability of the calculated measure insofar as (1) sampling fluctuations and (2) variable errors of measurement are concerned.

2. Standard and probable errors do not reveal nor measure (1) the effect upon any statistical measure of constant errors, or errors of bias; nor do they tell us (2) when our sample is not representative of its parent population.

PROBLEMS

1. Given: $M = 26.40$; $\sigma = 3.20$; $N = 100$.
 - (a) What are the chances that the true mean of the population from which the given 100 cases is a random sample will be greater than 27?
 - (b) What are the chances that the true mean lies between 26 and 27?
 - (c) What are the chances that the σ of the population lies between 3.10 and 3.30? That it is greater than 3.50?
2. Given: $Mdn = 72.40$; $Q = 12.84$; $N = 81$.
 - (a) What are the chances that the true median of the population from which this sample is drawn is above 75?
 - (b) That it lies between 70 and 74?
 - (c) What are the chances that the true Q is not greater than 15?
 - (d) That it lies between 10 and 14?
3. The mean of a distribution is K and the PE_M is 2.50. What are the chances that the true mean will not diverge from the obtained by more than (a) 1.00; (b) 3.00; (c) 10.00?
4. The difference between two medians ($Mdn_1 - Mdn_2$) is 3.60, and the $PE_D = 3.00$.
 - (a) What are the chances that the obtained difference is significant?
 - (b) That it is 1.00 or more?
 - (c) What percent is the obtained difference of the difference necessary for complete reliability?
5. In the first trial of a practice period, twenty-five 12 year olds have a mean score of 80.00 and a σ of 8.00 upon a digit-symbol learning test. On the tenth trial, the mean is 88.00 and the σ is 10.00. The r between scores on the first and tenth trials is .40.
 - (a) Is the gain in score significant?
 - (b) Is the increase in variability significant?
6. Two groups of high school pupils are matched for general intelligence and for initial ability in a biology test. Group 1 is taught by the lecture method, and Group 2 by the lecture-demonstration method. At the end of the term, results are as follows:

	Group 1 (control)	Group 2 (experimental)
N	60	60
Mean general intelligence score	135.50	136.35
Mean initial score on the biology test	42.30	42.50
σ of initial scores on the biology test	5.36	5.38
Mean final score on the biology test	54.54	56.74
σ of final scores on the biology test	6.34	7.25
r (between final scores on the biology test) =	.50	

- (a) Is the difference between the final scores made by Groups 1 and 2 upon the biology test significant?
 - (b) Is the difference in the variability of the final scores made by Groups 1 and 2 significant?
7. Calculate measures of skewness and kurtosis for each of the four distributions in Chapter II, problem 1, page 30. Compute also standard errors of Sk and Ku by the formulas given on pages 229 and 230. Determine whether any of these distributions departs significantly from the normal form.
8. In a city high school of 5000 pupils, 52.3% are girls and 47.7% are boys. This school is one of several of approximately the same size. If it is a good sample of the high school population in this city, is there a significant difference between the percentages of boys and girls enrolled in the high schools?
9. In an institution, 80 delinquent and 80 non-delinquent boys of the same age, same I Q, and roughly the same social status furnish the following data:
 - (a) 40% of the delinquent, and 20% of the non-delinquent come from "poor" homes.
 - (b) 74% of the delinquent and 44% of the non-delinquent score above the "normal" median on a neurotic inventory.
 - (c) 65% of the delinquent and 50% of the non-delinquent cheat on a certain test.
 Are any of these differences significant?
10. In a random sample of 100 cases each from four groups A, B, C, and D, the following results were obtained:

	A	B	C	D
Mean	101.00	104.00	93.00	86.00
σ	10.00	11.00	9.60	8.50

What are the chances that, in general, the mean of

- (a) the A's is better than the mean of the B's.
- (b) the A's is 5 better than the mean of the C's.
- (c) the A's is 10 better than the mean of the D's.

What are the chances that

- (a) a B will be better than the mean A.
- (b) a B will be better than the mean C.
- (c) a B will be better than the mean D.

ANSWERS

1. (a) 3 in 100
(b) 86 in 100
(c) 34 in 100; 10 in 100
2. (a) 16 in 100 (c) 90 in 100
(b) 55 in 100 (d) 71 in 100
3. (a) 21 in 100 (b) 58 in 100 (c) 99 in 100
4. (a) 79 in 100
(b) 72 in 100
(c) 30%
5. (a) Yes. $D/\sigma_D = 4$
(b) No. $D/\sigma_{D\sigma} = 1.20$
6. (a) $D/\sigma_D = 2.49$; 99 chances in 100 of a significant difference.
(b) $D/\sigma_{D\sigma} = 1.20$; 88 " " " " " "
7. Distribution Sk/σ_{sk} Ku/σ_{ku} Deviation from normality not significant

1	— .23	.55	"	"	"	"	"
2	.51	— .38	"	"	"	"	"
3	.33	.93	"	"	"	"	"
4	— .13	.68	"	"	"	"	"
8. Yes. $D/\sigma_{Dp} = 4.6$

9. (a) $D/\sigma_{D_p} = 2.83$ (998 chances in 1000). Probably significant
(b) $D/\sigma_{D_p} = 4.05$. Significant
. (c) $D/\sigma_{D_p} = 1.94$ (97 chances in 100). Not significant
10. (a) 2 in 100
* (b) 98 in 100
(c) 100%
(a) 61 in 100
(b) 84 in 100
(c) 95 in 100

CHAPTER IX

LINEAR CORRELATION

I. WHAT IS MEANT BY CORRELATION

IN previous chapters, we have been chiefly concerned with methods of computing statistical measures designed to represent in a reliable way the performance of an individual or a group in some defined capacity or trait. Frequently, however, it is of more importance to examine the *relationship* of one ability to another than it is to measure performance in either trait alone. Are certain abilities closely related, and others relatively independent? Is it true, for example, that good pitch discrimination accompanies musical achievement; or that bright children tend to be less neurotic than average children? If we know the general intelligence of a child, as measured by a standard test, can we say anything about his probable scholastic achievement as represented by grades? Problems like these and many others which involve the relations among abilities are studied by the method of correlation.

When the relationship between two sets of measures is "linear," i.e., can be described by a straight line,* the correlation between the scores may be expressed by the "product-moment" coefficient of correlation. This coefficient is designated by the letter r . The method of calculating r will be outlined in Section III. Before taking up the details of calculation, we shall try to make clear what correlation means, and how r measures relationship.

Let us consider, first, a situation in which relationship is fixed and unchanging. The circumference of a circle is always

* See pages 289-291 for a fuller discussion of "linear" relationship.

3.1416 times its diameter ($C = 3.1416D$), and this equation holds no matter how large or how small the circle, or in what part of the world we find it. Each time the diameter of a circle is increased or decreased, the circumference is increased or decreased by just 3.1416 times the same amount. In short, the dependence of circumference upon diameter is complete; hence, the correlation between the two dimensions is said to be perfect, and $r = 1.00$. In the same way, the relationship between two abilities, as represented by two sets of measures, may also be perfect. Suppose, for example, that 100 students have exactly the same standing in two tests — the student who ranks first in the one test ranks first in the other, the student who ranks second in the first test ranks second in the other, and this one-to-one correspondence holds throughout the entire list. The relationship here is perfect since the relative position of each man is exactly the same in one test as in the other, and the coefficient of correlation is 1.00.

Now let us consider the case in which there is *no* correlation present. Suppose that we have administered to 100 college seniors the Army Alpha Examination and a simple "tapping test" in which the number of separate taps made in 30 seconds is recorded. Let the mean Alpha score for the whole group be 175, and the mean tapping rate be 185 taps in 30 seconds. Now suppose that when we divide our group into three sub-groups in accordance with the size of their Alpha scores, we find that the mean tapping rate of the superior or "high" group (whose mean Alpha score is 190) is 184 taps in 30"; the mean tapping rate of the "middle" group (whose mean Alpha score is 175) is 186 taps in 30"; and the mean tapping rate of the lowest group (whose mean Alpha score is 160) is 185 taps in 30". Since the tapping rate is almost identically the same for all three groups, it is clear that from a student's tapping rate alone we should be unable to draw any conclusion as to his probable performance upon Alpha. A tapping rate of 185 is as liable to be found with an Army Alpha score of 150, as with one of 175 or even 200. In other words, there is no correspond-

ence between the sizes of the scores made by the members of our group upon the two tests; and hence, r , the coefficient of correlation, is zero.

Perfect relationship, then, is expressed by a coefficient of 1.00, and just no relationship by a coefficient of .00. Between these two limits, varying degrees of relation are indicated by such coefficients as .33, or .65, or .92. A coefficient of correlation falling between .00 and 1.00 always implies *some* degree of positive association, the degree of the association depending upon the size of the coefficient.

Relationship may be negative as well as positive; that is, a *large* degree of one trait may be associated with a *small* degree of another. When negative or inverse relationship is perfect, $r = -1.00$. To illustrate, suppose that in a small class of 10 school boys, the boy who stands first in Latin ranks lowest (10th) in shop work; the boy who stands second in Latin ranks next to the bottom (9th) in shop work; and that each boy stands just as far from the top of the list in Latin as from the bottom of the list in shop work. Here the correspondence between achievement in Latin and shop work is one-to-one and hence definite enough, but the direction of the relationship is inverse (or reverse) and $r = -1.00$. Negative coefficients may range from -1.00 up to .00, just as positive coefficients may range from .00 up to 1.00. Coefficients of $-.20$, $-.50$, or $-.80$ indicate increasing degrees of negative or inverse relationship, just as positive coefficients of .20, .50, and .80 indicate increasing degrees of positive relationship.

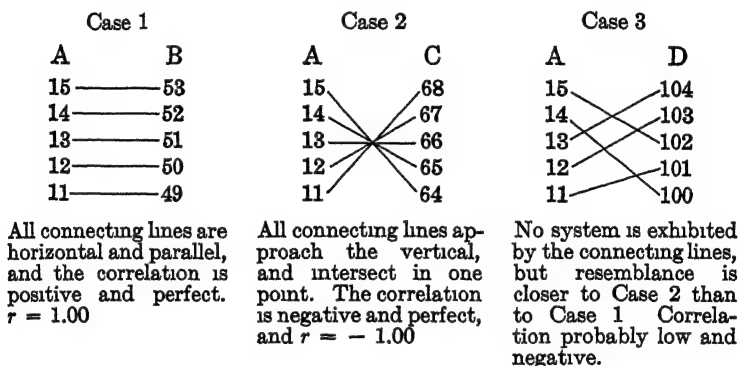
The notion underlying correlation can often be most readily comprehended from a simple graphic treatment. Three examples will be given to illustrate values of r of 1.00, -1.00 , and approximately 00. Correlation is rarely computed when the number of cases is less than 25, so that the examples here presented must be considered to have illustrative value only.

Suppose that four tests, A, B, C, and D, have been administered to a group of five children. The children have been arranged in order of merit on Test A which is then compared

separately with Tests B, C, and D (not in order of merit) to give the following three cases:

Case 1			Case 2			Case 3		
Pupil	A	B	Pupil	A	C	Pupil	A	D
a	15	53	a	15	64	a	15	102
b	14	52	b	14	65	b	14	100
c	13	51	c	13	66	c	13	104
d	12	50	d	12	67	d	12	103
e	11	49	e	11	68	e	11	101

Now if the *second* series of scores under each case (i.e., B, C, and D) is arranged in order of merit from the highest score down, and the two scores earned by each child are connected by a straight line, we have the following graphs:



The more nearly the lines connecting the paired scores are horizontal and parallel, the higher the positive correlation. The more nearly the connecting lines are vertical and tend to intersect in one point, the lower and more negative the correlation. When the connecting lines show no systematic trend, the correlation approaches zero.

To summarize our discussion up to this point, coefficients of correlation range over a scale which extends from -1.00 through $.00$ to 1.00 . A positive correlation indicates that *large* amounts of the one variable tend to accompany *large* amounts of the other; a negative correlation indicates that *small* amounts

of the one variable tend to accompany *large* amounts of the other. A zero correlation indicates no consistent relationship. We have illustrated above only perfect positive, perfect negative, and approximately zero correlation in order to bring out the meaning of correlation in a striking way. Only rarely if ever, however, will a coefficient fall at either extreme of the scale, i.e., at 1.00 or -1.00. In most actual problems calculated r 's fall at intermediate points, such as .72, -.26, .50, etc. Such r 's are to be interpreted as "high" or "low" depending in general upon how close they are to ± 1.00 . Interpretation of the degree of relationship expressed by r in terms of various criteria will be discussed later on pages 342-356.

II. THE COEFFICIENT OF CORRELATION

1. The Coefficient of Correlation as a Ratio

The product-moment coefficient of correlation may be thought of essentially as that ratio which expresses the extent to which changes in one variable are accompanied by — or are dependent upon — changes in a second variable. As an illustration, consider the following simple example which gives the paired heights and weights of five college seniors:

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Student	Ht. in inches	Wt. in lbs.						
	X	Y	x	y	xy	$\frac{x}{\sigma_x}$	$\frac{y}{\sigma_y}$	$\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}$
a	72	170	3	0	0	1.5	0	0
b	69	165	0	-5	0	0	-.41	0
c	66	150	-3	-20	60	-1.5	-1.63	2.5
d	70	180	1	10	10	.5	.82	.4
e	68	185	-1	15	-15	-.5	1.23	-.6
					55			2.3

$$M_X = 69 \text{ in. } \sigma_x = 2 \text{ in.}$$

$$M_Y = 170 \text{ lbs. } \sigma_y = 12.25 \text{ lbs.}$$

$$\text{correlation} = \frac{\sum \left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y} \right)}{N} = \frac{2.3}{5} = .46$$

From the X and Y columns it is evident that tall students tend to be somewhat heavier than short students, and hence the correlation between height and weight is almost certainly posi-

tive. The mean height is 69 inches, the mean weight 170 pounds, and the σ 's are 2 inches and 12.25 pounds, respectively. In column 4 are given the deviations (x 's) of each man's height from the mean height, and in column 5 the deviations (y 's) of each man's weight from the mean weight. The product of these paired deviations (xy 's) is a measure of the agreement between individual heights and weights, and the higher the sum of the xy column the greater the degree of correspondence. When agreement is perfect (and $r = 1.00$), the Σxy column has its maximum value. It may be surmised — and with much reason — that the sum of the xy 's divided by N (i.e., $\frac{55}{5} = 11$) should give a suitable measure of the relationship between X and Y . Such an average is not a stable measure of relationship, however, as it depends directly upon the *units* in which height and weight have been expressed, and consequently will vary (as shown in the example below) if centimeters and kilograms, say, are employed instead of inches and pounds. One may avoid the troublesome matter of differences in units by first dividing each x and each y by its own σ , i.e., by expressing each deviation as a standard or z -score. The sum of the products of the standard scores (column 9) divided by N will then yield a ratio which is a stable expression of relationship. This ratio is the "product-moment" * coefficient of correlation. Its value of .46 indicates a fairly high positive correlation between height and weight in this small sample. The student should note that our ratio or correlation coefficient is simply the *average product* of the standard scores of corresponding X and Y measures.

Let us now investigate the effect upon our ratio of changing the units in terms of which X and Y have been expressed. In the example below, the heights and weights of the same five students are expressed (to the nearest whole number) in centimeters and kilograms instead of in inches and pounds:

* The sum of the deviations from the mean (raised to some power) and divided by N is called a "moment." When pairs of deviations in x and y are multiplied together, summed, and divided by N (to give $\frac{\Sigma xy}{N}$) the term "product-moment" is used.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Student	Ht. in cms	Wt. in kgs.						
	X	Y	x	y	xy	$\frac{x}{\sigma_x}$	$\frac{y}{\sigma_y}$	$\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}$
a	183	77	8	0	0	1.6	0	0
b	175	75	0	-2	0	0	.36	0
c	168	68	-7	-9	63	-1.4	-1.61	2.3
d	178	82	3	5	15	.6	.89	.5
e	173	84	-2	7	-14	-.4	1.25	-.5
					64			2.3

$$M_X = 175 \text{ cms.} \quad \sigma_x = 5 \text{ cms.} \\ M_Y = 77 \text{ kgs.} \quad \sigma_y = 5.6 \text{ kgs.} \quad \text{correlation} = \frac{\sum \left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y} \right)}{N} = \frac{2.3}{5} = .46$$

The mean height of our group is now 175 cms. and the mean weight 77 kgs.; the σ 's are 5 cms. and 5.6 kgs., respectively. Note that the sum of the xy column, namely, 64, differs by 9 from the sum of the xy 's in the example above in which inches and pounds were the units of measurement. However, when deviations are expressed as standard scores, the sum of their products $\left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y} \right)$ divided by N gives a ratio of .46, as before.

The ratio

$$\frac{\sum \left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y} \right)}{N}$$

is a measure of relationship which remains constant no matter in what units X and Y are expressed. When this ratio is written $\frac{\sum xy}{N\sigma_x\sigma_y}$ it becomes the well-known expression for r , the product-moment coefficient of correlation, originally developed by Karl Pearson.* The application of this formula to correlation problems will be made in Section III following.

* The coefficient of correlation, r , is often called the "Pearson r " after Professor Karl Pearson who developed the product-moment method, following the earlier work of Galton and Bravais. See Walker, H. M., *Studies in the History of Statistical Method*, 1929, Chapter 5, pp. 96-111.

2. The Scatter Diagram and the Correlation Table

When N is small, the ratio method described in the preceding section is often employed to compute directly the coefficient of correlation between two sets of data (p. 255). When N is large, however, much time and labor may be saved by first arranging the data in the form of a diagram or chart, and calculating our deviations from assumed, instead of from actual, means. Let us consider the diagram in Figure 43. This chart, which is called a "scatter diagram" or "scattergram," represents the paired heights and weights of 120 college students. The construction of a scattergram is a relatively simple matter. Along the left hand margin from bottom to top are laid off the steps of the height distribution, measurement expressed in inches; and along the top of the diagram from left to right are laid off the steps of the weight distribution, measurement expressed in pounds. Each of the 120 men is represented on the diagram with respect to both height and weight. Suppose that a man weighs 150 lbs. and is 69 in. tall. His weight locates him in the sixth column from the left, and his height in the third row from the top. Accordingly, a "tally" is placed in the third cell of the sixth column. There are three tallies in all in this cell, that is, there are three men who weigh from 150 to 159 lbs., and are 68-69 in. tall. Each of the 120 men is represented by a tally in a cell or square of the table in accordance with the two characteristics, height and weight. Along the bottom of the diagram in the f_x row is tabulated the number of men who fall in each weight-interval; while along the right hand margin in the f_y column is tabulated the number of men who fall in each height-interval. Both the f_y column and f_x row must total 120, the number of men in all. After all of the tallies have been made, the frequency in each cell is totaled and entered on the diagram. The scattergram is then a *correlation table*.

Several interesting facts may be gleaned from this correlation table as it stands. For example, all of the men of a given weight-interval may be studied with respect to distribution of height. In the third column there are 28 men all of whom

		Weight in Pounds (X-Variable)									
		100-109	110-119	120-129	130-139	140-149	150-159	160-169	170-179	f_y	M_{wt}
Height in Inches (Y-Variable)	72-73								1	1	174.5
	70-71			1	3	3	4	2	3	16	152.0
	68-69			4	11	6	3	2	2	28	142.4
	66-67		2	9	11	8	2	1		33	135.1
	64-65	1	5	7	10	3				26	128.0
	62-63	1	2	7	1	2				13	125.3
	60-61	1	1		1					3	117.8
		f_x	3	10	28	37	22	9	5	6	120
		M_{ht}	62.5	64.1	65.4	66.6	67.0	68.9	68.9	70.2	

(1)

(2)

Weight	Mean ht. for given wt. interval	Height	Mean wt. for given ht. interval
170-179	70.2	72-73	174.5
160-169	68.9	70-71	152.0
150-159	68.9	68-69	142.4
140-149	67.0	66-67	135.1
130-139	66.6	64-65	128.0
120-129	65.4	62-63	125.3
110-119	64.1	60-61	117.8
100-109	62.5		

FIG. 43. A Scattergram and Correlation Table Showing the Paired Heights and Weights of 120 Students.

weigh 120-129 lbs. One of the 28 is 70-71 in. tall; four are 68-69 in. tall; nine are 66-67 in. tall; seven are 64-65 in. tall; and seven are 62-63 in. tall. In the same way, we may classify all of the men of a given height-interval with respect to weight distribution. Thus, in the row next to the bottom, there are 13 men all of whom are 62-63 in. tall. Of this group one weighs 100-109 lbs.; two weigh 110-119 lbs.; seven weigh 120-129

lbs.; one weighs 130–139 lbs.; and two weigh 140–149 lbs. It is fairly clear that the “drift” of paired heights and weights is from the upper right hand section of the diagram to the lower left hand section. Even a superficial examination of the diagram, therefore, reveals a fairly marked tendency for heavy medium, and light men to be tall, medium, and short, respectively; and this general relationship holds in spite of the scatter of heights and weights within any given “array” (an array is the distribution of cases within a given column or row). Even before making any calculations, then, we should probably be willing to estimate the correlation between height and weight to be positive and fairly high.

Let us now go a step further and calculate the mean height of the three men who weigh 100–109 lbs., the men in column 1. The mean height of this group (using the assumed mean method described in Chapter II, p. 26) is 62.5 inches, and this figure has been written in at the bottom of the correlation table. In the same way, the mean heights of the men who fall in each of the succeeding weight-columns have been written in at the bottom of the diagram. These data have been tabulated in a somewhat more convenient form in (1) below the diagram. From this summary, it appears that an actual weight increase of approximately 80 lbs. (180–100) corresponds to an increase in mean height of 7.7 inches. In this group of 120 students therefore, an increase of 80 lbs. as we go from the lightest to the heaviest man is paralleled by an *average* increase of 7.7 inches in height. It seems clear that the correlation between height and weight is positive.

Let us now shift from height to weight, and applying the method used above, find the change in *mean weight* which corresponds to the given change in height. The mean weight of the three men in the bottom row of the diagram is 117.8 lbs. The mean weight of the 13 men in the next row from the bottom (who are 62–63 inches tall) is 125.3 lbs. The mean weights of the men who fall in the other rows have been written in their appropriate places in the M_{wt} column. In the summary of

results under (2), we find that in this group of 120 men an increase of about 14 inches in height is accompanied by an average increase of about 56.7 lbs. in weight. Thus it appears that the taller the man the heavier he tends to be, and again the correlation between height and weight is seen to be positive. In Section III we shall show how the coefficient of correlation may be calculated from a correlation table. We are concerned here only with the principles of construction of the scatter diagram and the correlation table.

3. The Graphic Representation of the Correlation Coefficient

It is often helpful in understanding how the correlation coefficient measures relationship if the student can see how a correlation of .00 or .50, say, looks graphically. Figure 44 (1) pictures a correlation of .50. The data in the table are artificial, and were selected so as to bring out the relationship in as unequivocal a fashion as possible. The scores laid off along the top of the correlation table from left to right will be referred to simply as the *X*-test "scores," and the scores laid off at the left of the table from bottom to top as the *Y*-test "scores." As was done in Figure 43, the means of each *Y*-row have been calculated and are entered on the chart opposite the appropriate *Y*-rows; and the means of the *X*-columns have been entered at the bottom of each *X*-column.

The means of each *Y*-array, that is, the means of the "scores" falling in each *X*-column, are indicated on the chart by small crosses. Through these crosses a line, called a *regression* line,* has been drawn. This line represents the change in the *mean* value of *Y* over the given range of *X*. The same facts are shown by the change in calculated "mean heights" in Figure 43. In similar fashion, the means of each *X*-array, i.e., the means of the scores in each *Y*-row, are designated on the chart by small circles, through which another line has been drawn. This second *regression* line shows the change in the *mean* value of *X*

* Regression lines have important properties; they will be defined and discussed more fully in Chapter X.

over the given range of Y . These two lines together represent the "linear" or straight line relationship between X and Y .

The closeness of association or degree of correspondence between the X - and Y -tests is indicated by the relative position

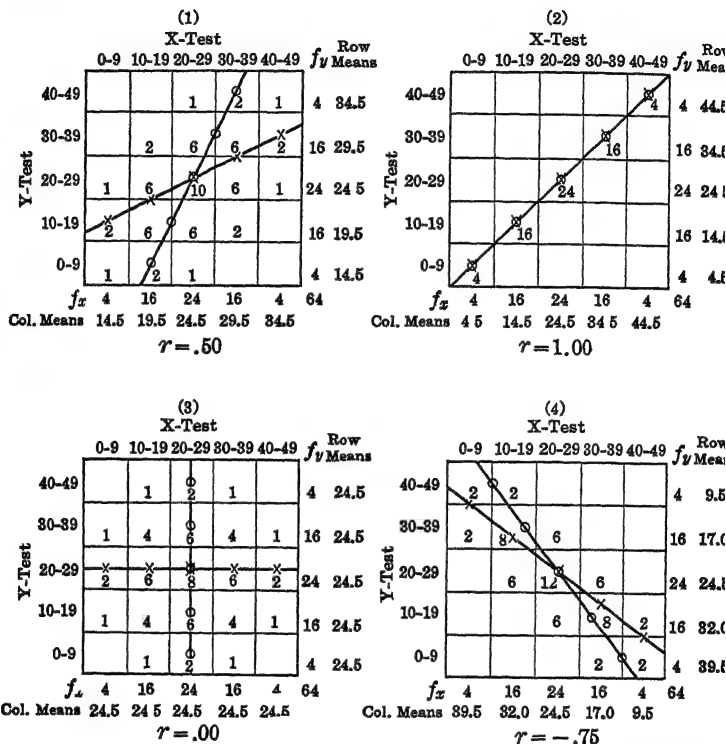


FIG. 44. The Graphical Representation of the Correlation Coefficient.

of these two regression lines. When the correlation is positive and perfect, the two regression lines close up like a pair of scissors to form one line. Chart (2) in Figure 44 shows how the two regression lines look when $r = 1.00$, and the correlation is perfect. Note that the entries in Chart (2) are concentrated along the diagonal from the upper right to the lower left hand

section of the diagram. There is no "scatter" of scores in the successive columns or rows, all of the scores in any array being concentrated within one cell. If Chart (2) represented a correlation table of height and weight, we should know that the tallest man was the heaviest, the next tallest man the next heaviest, and that throughout the group the correspondence of height and weight was 1:1.

A very different picture from that of perfect correlation is presented in Chart (3) where the correlation is .00. Here the two regression lines through the means of the columns and rows have spread out until they are perpendicular to each other. There is no change in the *mean Y*-score over the whole range of *X*, and no change in the *mean X*-score over the whole range of *Y*. This is analogous to the situation described on page 252, in which the mean tapping rate of a group of students was the same for those with "high," "middle," and "low" Army Alpha scores. When the correlation is zero, there is no way of telling from a subject's performance in one test what his performance will be in the other test. The best one can do is to select the mean as the "most probable" score for every one.

Chart (4) in Figure 44 represents a correlation coefficient of $-.75$. Negative relationship is shown by the fact that the regression lines through the means of the columns and rows run from the upper left to the lower right hand section of the diagram. The regression lines are closer together than in Chart (1) where the correlation is .50, but are still separated. If this chart represented a correlation table of height and weight, we should know that the tendency was strong for tall men to be light, and for short men to be heavy.

The charts in Figure 44 represent, as was stated above, the linear relationship between sets of artificial test scores. The data were selected so as to be symmetrical around the means of each column and row, and hence the regression lines go through *all* of the crosses or *through all* of the circles in the successive columns and rows. It is rarely if ever true, however, that the regression lines pass through *all* of the means of the columns

and rows in a correlation table which represents actual test scores or other real measures. Figure 45 which reproduces the correlation table of heights and weights given on page 256 illustrates this fact. The mean heights of the men in the weight (X) columns are indicated by crosses, and the mean weights of the men in the height (Y) rows by circles, as in Figure 44. Note that the series of short lines joining the suc-

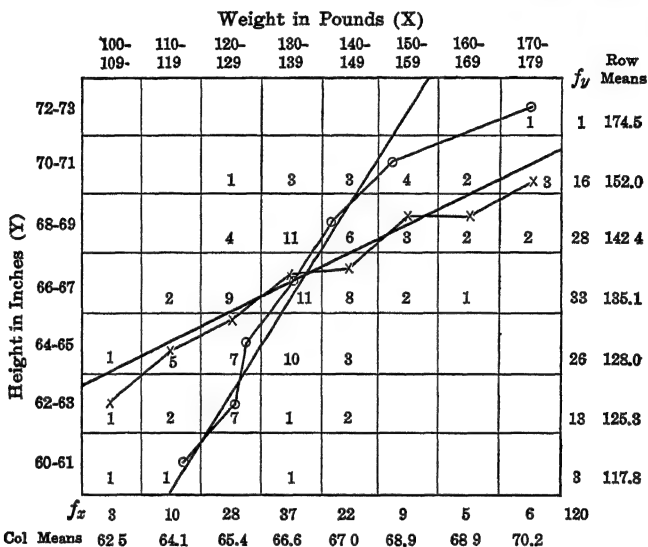


FIG. 45. Graphical Representation of the Correlation between Height and Weight in a Group of 120 College Students (Fig. 43).

cessive crosses or circles give a decidedly jagged appearance. Two straight lines have been drawn in to describe the general trend of these irregular lines. These two lines go through, or as close as possible to, the crosses or the circles, more consideration being taken of those points near the middle of the chart (because they are based upon more data) than of those at the extremes (which are based upon few scores). Regression lines when their equations fulfill certain mathematical criteria to be given later (p. 291) are called lines of "best fit." Such lines

describe better than any other *straight* lines the "run" or "drift" of the crosses and circles across the chart.

In Chapter X we shall develop the equations for these "best fitting" lines and show how they may be drawn in to describe the trend of irregular points on any correlation table. For the present, the important fact to get clearly in mind is that when correlation is "linear," the course of the means of the columns and rows in a correlation table can be described by two straight lines and that the closer together these two lines are the higher is the correlation.

III. THE CALCULATION OF THE COEFFICIENT OF CORRELATION BY THE PRODUCT-MOMENT METHOD

1. The Calculation of r from a Correlation Table

Having discussed briefly the meaning of correlation in the last sections, we shall now proceed to the calculation of the coefficient of correlation by the "product-moment" method. Figure 46 will serve as an illustration of the computations involved. This correlation table gives the paired heights and weights of 120 college students, and is derived from the scattergram for the same data shown in Figure 43. The following outline of the steps in the process of calculating r will be best understood if the student will constantly refer to Figure 46 as he reads through each step.

Step 1

Construct a scattergram for the two tests to be correlated, and from it draw up a correlation table as described on page 258.

Step 2

The distribution of heights for the 120 men is in the f_y column at the right of the diagram. Assume a mean for the height distribution, using the rules given in Chapter II, page 26, and draw double lines to mark off the *row* which contains the assumed mean (ht). The mean for the height distribution has been assumed at 66.5 in. (midpoint of interval 66-67) and

the y 's* have been taken from this point. Now fill in the fy' and fy'^2 columns. From the first column c_y , the correction in units of step, is obtained; and this correction together with the sum of the fy'^2 will give the σ of the height distribution, σ_y . As shown by the calculations in Figure 46, the value of σ_y is 2.62 inches.

The distribution of the weights of the 120 men is in the f_x row at the bottom of the diagram. Assume a mean for the weight distribution, and draw double lines to designate the column which contains the assumed mean (wt). Since the mean for the weight distribution was taken at 134.5 pounds (mid-point of interval 130-139), the x 's are taken from this point. Fill in the fx' and the fx'^2 rows; from the first calculate c_x , the correction in units of step, and from the second calculate σ_x , the σ of the entire weight distribution. In Figure 46, the value of σ_x is found to be 15.54 pounds.

Step 3

The calculations in Step 2 simply repeat the now familiar process of calculating σ by the Assumed Mean method. Our first new task is to fill in the $\Sigma x'y'$ column at the right of the chart. Since the entries in this column may be either + or -, two columns are provided under $\Sigma x'y'$. Calculation of the entries in the $\Sigma x'y'$ column may be illustrated by considering, first, the single entry in the only occupied cell in the topmost row. The deviation of this cell from the AM of the weight distribution, that is, its x' , is 4 steps, and its deviation from the AM of the height distribution, that is, its y' , is 3 steps. Hence, the product of the deviations of this cell from the two AM 's is 4×3 or 12; and a small figure 12 is placed in the upper right hand corner of the cell† The "product-deviation" of the one

* The prime (') of the x 's and y 's indicates that these deviations are taken from the assumed means of the X and Y distributions (see p. 27).

† We may consider the coordinates of this cell to be $x' = 4$, $y' = 3$. The x' is obtained by counting over four steps from the vertical column containing the AM (wt), and y' by counting up three steps from the horizontal row containing the AM (ht). The unit of measurement is the step-interval.

[illegible]

entry in this cell is 1 (4×3) or 12 also, and a figure 12 is placed in the lower left hand corner of the cell. This figure shows the product of the deviations of this single entry from the AM 's of the two distributions. Since there are no other entries in the cells of this row, 12 is placed at once under the $+$ sign in the $\Sigma x'y'$ column.

Consider, now, the next row from the top, taking the cells in order from right to left. The cell just below the one whose product-deviation we have just found also deviates 4 steps from the AM (wt) (its x' is 4), but its deviation from the AM (ht) is only 2 steps (its y' is 2). The product-deviation of this cell, therefore, is 4×2 or 8, as shown by the small figure (8) in the upper right hand corner of the cell. There are 3 entries in this cell, and since each has a product-deviation of 8, the final entry in the lower left hand corner of the cell is $3(4 \times 2)$ or 24. The product-deviation of the second cell in this row is 6 (its x' is 3 and its y' is 2) and since there are 2 entries in the cell, the final entry is $2(3 \times 2)$ or 12. Each of the 4 entries in the third cell over has a product-deviation of 4 (since $x' = 2$ and $y' = 2$), and hence the final cell entry is 16. In the fourth cell, each of the 3 entries has a product-deviation of 2 ($x' = 1$, and $y' = 2$) and the final entry is 6. The final entry in the fifth cell over, the cell in the AM (wt) column, is 0, since x' is 0, and accordingly $3(2 \times 0)$ must be 0. Note carefully the final entry (-2) in the last cell of the row. Since the deviations of this cell are $x' = -1$, and $y' = 2$, the product $1(-1 \times 2) = -2$, and the final entry is negative. Now we may total up the plus and minus entries in this row and enter the results, 58 and -2 , in the $\Sigma x'y'$ column under the appropriate signs.

The final entries in the cells for the other rows of the table and the sums of the product-deviations of each row are obtained as illustrated for the two rows above. The student should bear in mind in calculating $x'y'$'s that the product-deviations of *all* entries in the cells in the *first* and *third* quadrants of the table are positive, while the product-deviations of *all* entries

in the *second* and *fourth* quadrants are negative (p. 63). It should be remembered, too, that all entries either in the column containing the AM_x or the row containing the AM_y have zero product-deviations, since in the one case the x' and in the other the y' equals zero.

Since all entries in a given row have the same y' , the arithmetic of calculating $x'y'$'s may often be considerably reduced if each entry in a row-cell is first multiplied by its x' , and the sum of these deviations ($\Sigma x'$) multiplied once for all by the common y' — the y' of the row. The last two columns $\Sigma x'$ and $\Sigma x'y'$ contain the entries for the rows. To illustrate the method of calculation, in the second row from the bottom, taking the cells in order from right to left, upon multiplying the entry in each cell by its x' , we have $(2 \times 1) + (1 \times 0) + (7 \times -1) + (2 \times -2) + (1 \times -3)$ or -12 . If we multiply this "deviation-sum" by the y' of the whole row (i.e., by -2) the result is 24 which is the final entry in the $\Sigma x'y'$ column. Note that this entry checks the 28 and -4 entered separately in the $\Sigma x'y'$ column by the longer method. This shorter method is often employed in printed correlation charts and is recommended for use as soon as the student understands fully how the cell entries are obtained.

The $\Sigma x'y'$ may be checked by computing the product-deviations and summing for columns instead of rows. The two rows at the bottom of the diagram, $\Sigma y'$ and $\Sigma x'y'$, show how this is done. We may illustrate with the first column on the left, taking the cells from top to bottom. Multiplying the entry in each cell by its appropriate y' , we have $(1 \times -1) + (1 \times -2) + (1 \times -3)$ or -6 . When this entry in the $\Sigma y'$ row is multiplied by the common x' of the column (i.e., by -3) the final entry in the $\Sigma x'y'$ row is 18 . The sum of the $x'y'$ computed from the rows should check the sum of the $x'y'$ computed from the columns.

Two other useful checks are shown in Figure 46. The fy' will equal the $\Sigma y'$ and the fx' will equal the $\Sigma x'$ if no error has been made. The fy' and the fx' are the same as the $\Sigma y'$ and

$\Sigma x'$; although these columns and rows are designated differently, they denote in each case the sum of deviations times frequency.

Step 4

When all of the entries in the $\Sigma x'y'$ column have been made, and the column totaled, the coefficient of correlation may be calculated by the formula

$$r = \frac{\frac{\Sigma x'y'}{N} - c_x c_y}{\sigma_x \sigma_y} \quad (49)$$

(coefficient of correlation when deviations are taken from the assumed means of the two distributions) *

Substituting 146 for $x'y'$; .02 for c_y ; .18 for c_x ; 1.31 for σ_y ; 1.55 for σ_x ; and 120 for N , r is found to be .60.

It is very important to remember that c_x , c_y , σ_x and σ_y are all left in *units of step-interval* in formula (49). This is done because all product-deviations ($x'y'$'s) are in step-units, and hence it is desirable to keep *all* of the terms in the formula in step-units. Leaving the corrections and the two σ 's in units of step-interval facilitates computation, and does not change the result (i.e., the value of the coefficient of correlation).

Many printed charts for use in calculating coefficients of correlation by the product-moment method are available. The following may be mentioned as examples:

1. *The C-D Hand Correlation Chart*, by Edward Cureton and J. W. Dunlap, published by the Psychological Corporation, New York City.
2. *Dvorak Correlation Chart*, by August Dvorak, published by Longmans, Green and Co., New York City.
3. *Otis Correlation Chart*, by Arthur Otis, published by the World Book Co., Yonkers, N. Y.

* This formula for r differs slightly from the ratio formula developed on page 257. The fact that deviations are taken from assumed rather than from actual means makes it necessary to correct $\Sigma x'y'$ by subtracting the product of the two corrections c_x and c_y .

4. *Ruch-Stoddard Correlation Chart*, by G. M. Ruch and G. D. Stoddard, published by the University Bookstore, University of Iowa, Iowa City, Iowa.
5. *Simplex Correlation Form*, by A. R. Lauer, published by the Educational Test Bureau, Minneapolis, Minn.
6. *Thurstone Correlation Data Sheet*, by L. L. Thurstone, published by C. H. Stoelting and Co., Chicago, Ill.

There are several charts intended especially for use with a calculating machine.* Most of the printed charts have one or more checks upon the arithmetical work.

2. The Calculation of r from Ungrouped Data

- (1) The Formula for r When Deviations are Taken from the Means of the Two Distributions X and Y

Formula (49) assumes that all x' and y' deviations have been taken from the two assumed means; and hence it is necessary to correct $\frac{\sum x'y'}{N}$ by the amount of the two corrections, c_x and c_y (p. 28). When deviations have been taken from the actual means of the two distributions, instead of from assumed means, no correction is needed, as both c_x and c_y are zero. Under these conditions, formula (49) becomes

$$r = \frac{\sum xy}{N\sigma_x\sigma_y} \quad (50)$$

(coefficient of correlation when deviations are taken from the means of the two distributions)

which is the ratio for measuring correlation developed on page 257. If we write $\sqrt{\frac{\sum x^2}{N}}$ for σ_x and $\sqrt{\frac{\sum y^2}{N}}$ for σ_y (p. 42), the N 's cancel and formula (50) becomes

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}} \quad (51)$$

(coefficient of correlation when deviations are taken from the means of the two distributions)

* For example, *C-D Machine Correlation Chart*, by E. E. Cureton and J. W. Dunlap, published by Macmillan Co., New York City.

in which x and y are deviations from the actual means as in (50) and $\sqrt{\Sigma x^2}$ and $\sqrt{\Sigma y^2}$ are the sums of the squared deviations in x and y taken from the two means.

When N is fairly large, so that the data can be grouped into a correlation table, formula (49) is always used in preference to formulas (50) or (51) as it entails much less calculation. Formulas (50) and (51) may be used to good advantage, however, in finding the correlation between short, ungrouped, series (say, 25 cases or less). It is not necessary to tabulate the scores into a frequency distribution. An illustration of the use of formula (51) is given in Table 36. The problem is to find the correlation between the scores made by twelve adults on two tests of "controlled association." The steps in computing r may be outlined as follows:

TABLE 36

TO ILLUSTRATE THE CALCULATION OF r FROM UNGROUPED SCORES
WHEN DEVIATIONS ARE TAKEN FROM THE MEANS OF THE SERIES

Subject	Test 1 Test 2		x	y	x^2	y^2	xy
	X	Y					
A	50	22	- 12.5	- 8.4	156.25	70.56	105 00
B	54	25	- 8.5	- 5.4	72.25	29.16	45 90
C	56	34	- 6.5	3.6	42.25	12.96	- 23.40
D	59	28	- 3.5	- 2.4	12.25	5.76	8 40
E	60	26	- 2.5	- 4.4	6.25	19.36	11 00
F	62	30	- .5	- 4	.25	.16	.20
G	61	32	- 1.5	1.6	2.25	2.56	- 2.40
H	65	30	2.5	- 4	6.25	.16	- 1.00
I	67	28	4.5	- 2.4	20.25	5.76	- 10.80
J	71	34	8.5	3.6	72.25	12.96	30.60
K	71	36	8.5	5.6	72.25	31.36	47.60
L	74	40	11.5	9.6	132.25	92.16	110.40
	750	365			595.00 (Σx^2)	282.92 (Σy^2)	321.50 (Σxy)

$$M_X = 62.5 \quad M_Y = 30.4$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} = \frac{321.50}{\sqrt{595 \times 282.92}} = .78 \quad (51)$$

Step 1

Find the mean of Test 1 (X) and the mean of Test 2 (Y). The means in Table 36 are 62.5 and 30.4, respectively.

Step 2

Find the deviation of each score on Test 1 from its mean, 62.5, and enter it in column x . Next find the deviation of each score in Test 2 from its mean, 30.4, and enter it in column y .

Step 3

Square all of the x -deviations and all of the y -deviations and enter these squares in columns x^2 and y^2 , respectively. Total these columns to obtain Σx^2 and Σy^2 .

Step 4

Multiply the x and y deviations by rows, and enter these products (with due regard for sign) in the xy column. Total the xy column to get Σxy .

Step 5

Substitute for Σxy (321.50), for Σx^2 (595) and for Σy^2 (282.92) in formula (51), as shown in Table 36, and solve for r .

While formula (51) is useful in calculating r directly from two ungrouped series of scores, it has the same disadvantage as the "long method" of calculating means and σ 's described in Chapter III. The deviations in x and y , when taken from the actual means, are usually decimals and the multiplication and squaring of these values is often a tedious task. For this reason (even when working with short ungrouped series) it is often easier to assume means, calculate deviations from these AM 's, and apply formula (49). The procedure here is illustrated in Table 37 with the same data given in Table 36. Note that the two means, M_x and M_y , are first calculated. The two corrections, c_x and c_y , are found by subtracting AM_x from M_x and AM_y from M_y . Since deviations are taken from assumed means, fractions are avoided; and the calculations of $\Sigma x'^2$, $\Sigma y'^2$, and $\Sigma x'y'$ are readily made. Substitution in formula (49) then gives r .

(2) The Calculation of r from Raw Scores, i.e., When Deviations are Taken from Zero

The calculation of r may often be carried out most readily — especially when a calculating machine is available — by means

TABLE 37

TO ILLUSTRATE THE CALCULATION OF r FROM UNGROUPED SCORES
WHEN DEVIATIONS ARE TAKEN FROM THE ASSUMED
MEANS OF THE SERIES

Subject	Test 1 X	Test 2 Y	x'	y'	x'^2	y'^2	$x'y'$
A	50	22	- 10	- 8	100	64	80
B	54	25	- 6	- 5	36	25	30
C	56	34	- 4	4	16	16	- 16
D	59	28	- 1	- 2	1	4	2
E	60	26	0	- 4	0	16	0
F	62	30	2	0	4	0	0
G	61	32	1	2	1	4	2
H	65	30	5	0	25	0	0
I	67	28	7	- 2	49	4	- 14
J	71	34	11	4	121	16	44
K	71	36	11	6	121	36	66
L	74	40	14	10	196	100	140
	750	365			670	285	334
					($\Sigma x'^2$)	($\Sigma y'^2$)	($\Sigma x'y'$)

$$AM_X = 60.0 \quad AM_Y = 30.0$$

$$M_X = 62.5 \quad M_Y = 30.4$$

$$c_x = 2.5 \quad c_y = .4$$

$$c_x^2 = 6.25 \quad c_y^2 = .16$$

$$r = \frac{334}{12} - 1.00$$

$$r = \frac{7.04 \times 4.86}{7.04 \times 4.86} \quad (49)$$

$$\sigma_x = \sqrt{\frac{370}{12} - 6.25} = 7.04$$

$$\sigma_y = \sqrt{\frac{185}{12} - .16} = 4.86$$

$$r = .78$$

of the following formula which is based upon "raw" or obtained scores:

$$r = \frac{\Sigma XY - NM_X M_Y}{\sqrt{[\Sigma X^2 - NM_X^2][\Sigma Y^2 - NM_Y^2]}} \quad (52)$$

(coefficient of correlation calculated from raw or obtained scores)

In this formula, X and Y are obtained scores, and M_X and M_Y are the means of the X and Y series, respectively. ΣX^2 and ΣY^2 are the sums of the squared X and Y values, and N is the number of cases.

Formula (52) is derived directly from formula (49) by assuming the means of the X and Y tests to be zero. If AM_X and AM_Y are zero, each X and Y score is a deviation from its AM as it stands, and hence we work with the scores themselves. Since the correction, c , always equals $M - AM$, it follows that when

the AM equals 0, $c_x = M_x$, $c_y = M_y$ and $c_x c_y = M_x M_y$. Furthermore, when $c_x = M_x$ and $c_y = M_y$ and the "scores" are "deviations," the formula

$$\sigma_x = \sqrt{\frac{\sum f x'^2}{N} - c_x^2} \times \text{step}$$

(see p. 49) becomes

$$\sigma_x = \sqrt{\frac{\sum X^2}{N} - (M_x)^2}$$

and σ_y for the same reason equals $\sqrt{\frac{\sum Y^2}{N} - (M_y)^2}$. If we substitute these values for $c_x c_y$, σ_x , and σ_y in formula (49), the formula for r in terms of raw scores given in (52) is obtained.

An alternate form of (52) is often more useful in practice. This is

$$r = \frac{N \sum XY - \sum X \cdot \sum Y}{\sqrt{[N \sum X^2 - (\sum X)^2][N \sum Y^2 - (\sum Y)^2]}} \quad (53)$$

(coefficient of correlation calculated from raw or obtained scores)

This formula is obtained from (52) by substituting $\frac{\sum X}{N}$ for M_x , and $\frac{\sum Y}{N}$ for M_y in numerator and denominator, and canceling the N 's.

The calculation of r when original scores are taken as "deviations" from zero is shown in Table 38. The data are again the two sets of 12 scores obtained on the "controlled association" tests, the correlation for which was found to be .78 in Table 36. This short example is for the purpose of illustrating the arithmetic and must not be taken as a recommendation that formula (52) be used only with short series. As a matter of fact, formula (52) or (53) is most useful, perhaps, with long series, especially if one is working with a calculating machine.

The computation by formula (53) is straightforward and the method easy to follow, but the calculations become tedious if the scores are expressed in more than two digits. For this reason, when using formula (53) it will often greatly lessen the

TABLE 38

TO ILLUSTRATE THE CALCULATION OF r FROM UNGROUPED DATA
WHEN DEVIATIONS ARE ORIGINAL SCORES ($AM's = 0$)

Subject	Test 1 X	Test 2 Y	X^2	Y^2	XY
A	50	22	2500	484	1100
B	54	25	2916	625	1350
C	56	34	3136	1156	1904
D	59	28	3481	784	1652
E	60	26	3600	676	1560
F	62	30	3844	900	1860
G	61	32	3721	1024	1952
H	65	30	4225	900	1950
I	67	28	4489	784	1876
J	71	34	5041	1156	2414
K	71	36	5041	1296	2556
L	74	40	5476	1600	2960
	<u>750</u>	<u>365</u>	<u>47470</u>	<u>11385</u>	<u>23134</u>

$M_X = 62.50$
 $M_Y = 30.42$ (means to two decimals)

$$r = \frac{23134 - 12 \times 62.50 \times 30.42}{\sqrt{[47470 - 12 \times (62.50)^2][11385 - 12 \times (30.42)^2]}} \quad (52)$$

$$r = .78$$

arithmetical work, if we first "reduce" the original scores by subtracting constant quantities from each of the original X and Y scores. In Table 39, the same two series of 12 scores have been reduced by subtracting 65 from each of the X scores, and 25 from each of the Y scores. The reduced scores, entered in the table under X' and Y' , are first squared to give $\Sigma X'^2$ and $\Sigma Y'^2$, and then multiplied by rows to give $\Sigma X'Y'$. Substitution of these values in formula (53) gives the coefficient of correlation r . If the means of the two series are wanted, they may readily be found by adding to $\frac{\Sigma X'}{N}$ and $\frac{\Sigma Y'}{N}$ the amounts by which the X and Y scores were reduced (see computations in Table 39).

The method of computing r by first reducing the scores is usually superior to the method of applying formula (52) or (53) directly to the raw scores. This is because we deal with smaller whole numbers, and much of the arithmetic can be done mentally. When raw scores have more than two digits, they are

TABLE 39

TO ILLUSTRATE THE CALCULATION OF r FROM UNGROUPED DATA
 WHEN DEVIATIONS ARE ORIGINAL SCORES ($AM's = 0$)
 SCORES ARE "REDUCED" BY THE SUBTRACTION OF
 CONSTANTS FROM X AND Y

	Test 1	Test 2					
Sub- ject	X	Y	X'	Y'	X'^2	Y'^2	$X'Y'$
A	50	22	- 15	- 3	225	9	45
B	54	25	- 11	0	121	0	0
C	56	34	- 9	9	81	81	- 81
D	59	28	- 6	3	36	9	- 18
E	60	26	- 5	1	25	1	- 5
F	62	30	- 3	5	9	25	- 15
G	61	32	- 4	7	16	49	- 28
H	65	30	0	5	0	25	0
I	67	28	2	3	4	9	6
J	71	34	6	9	36	81	54
K	71	36	6	11	36	121	66
L	74	40	9	15	81	225	135
	750	365	- 30($\Sigma X'$)	65($\Sigma Y'$)	670($\Sigma X'^2$)	635($\Sigma Y'^2$)	159($\Sigma X'Y'$)

Subtract 65 from each X , and 25 from each Y to give X' and Y'

$$\begin{aligned}
 M_X &= \frac{\Sigma X'}{N} + 65 & M_Y &= \frac{\Sigma Y'}{N} + 25 \\
 &= \frac{-30}{12} + 65 & &= \frac{65}{12} + 25 \\
 &= 62.5 & &= 30.4
 \end{aligned}$$

$$\begin{aligned}
 r &= \frac{12 \times 159 + 30 \times 65}{\sqrt{[12 \times 670 - (-30)^2][12 \times 635 - (65)^2]}} & (53) \\
 &= \frac{3858}{4923} \\
 &= .78
 \end{aligned}$$

cumbersome to square and multiply unless reduced. The student should note that instead of 65 and 25 other constants might have been used to reduce the X and Y scores. If the smallest X and Y scores had been subtracted, namely, 50 and 22, all of the X' and Y' would, of course, have been positive. This is an advantage in calculation but these reduced scores would have been somewhat larger numerically than are the reduced scores in Table 39. In general, the best plan in reducing scores is to subtract constants which are close to the means.

The reduced scores are then both plus and minus, but are numerically about as small as we can make them.

(3) The Calculation of r by the Difference-Formula

It is apparent from the preceding sections that the product-moment formula for r may be written in several ways, depending upon whether deviations are taken from actual or assumed means, and upon whether raw scores or deviations are employed. The present section contributes still another formula for calculating r —namely, the difference-formula. This formula will complete our list of expressions for r , as it is believed that the student who understands the meaning and use of the correlation formulas given in this chapter will have no difficulty with other variations which he may encounter in the literature of psychology and education.*

The formula for r by the difference method is

$$r = \frac{\Sigma x^2 + \Sigma y^2 - \Sigma d^2}{2\sqrt{\Sigma x^2 \cdot \Sigma y^2}} \quad (54)$$

(coefficient of correlation by difference-formula, deviations from the means of the distributions)

in which $\Sigma d^2 = \Sigma (x - y)^2$.

The principal advantage of the difference-formula is that no cross products (xy 's) need be computed. For this reason, this formula is employed in many of the printed correlation charts. Formula (54) is illustrated in Table 40 with the same data used in Table 36 and elsewhere in this chapter. Note that the x , y , x^2 and y^2 columns repeat Table 36. The d or $(x - y)$ column is found by subtracting *algebraically* each y -deviation from its corresponding x -deviation. These differences are then squared and entered in the d^2 or $(x - y)^2$ column. Substitution of Σx^2 , Σy^2 , and Σd^2 in formula (54) gives $r = .78$.

The difference-formula may be employed when deviations have been taken from the assumed means of X and Y . The

* See the following article which lists 52 variations of the r -formula: Symonds, P. M., *Variations of the Product-Moment (Pearson) Coefficient of Correlation*, Journal of Educational Psychology, 1926, 17, pp 458-469.

TABLE 40

TO ILLUSTRATE THE CALCULATION OF r FROM UNGROUPED DATA BY
* THE DIFFERENCE-FORMULA, DEVIATIONS FROM THE MEANS

Subject	Test 1 Test 2		d						d^2
	X	Y	x	y	$(x - y)$	x^2	y^2	$(x - y)^2$	
A	50	22	-12.5	-8.4	-4.1	156.25	70.56	16.81	
B	54	25	-8.5	-5.4	-3.1	72.25	29.16	9.61	
C	56	34	-6.5	3.6	-10.1	42.25	12.96	102.01	
D	59	28	-3.5	-2.4	-1.1	12.25	5.76	1.21	
E	60	26	-2.5	-4.4	1.9	6.25	19.36	3.61	
F	62	30	-.5	-.4	-.1	.25	.16	.01	
G	61	32	-1.5	1.6	-3.1	2.25	2.56	9.61	
H	65	30	2.5	-.4	2.9	6.25	.16	8.41	
I	67	28	4.5	-2.4	6.9	20.25	5.76	47.61	
J	71	34	8.5	3.6	4.9	72.25	12.96	24.01	
K	71	36	8.5	5.6	2.9	72.25	31.36	8.41	
L	74	40	11.5	9.6	1.9	132.25	92.16	3.61	
						595.00	282.92	234.92	

$$M_X = 62.5$$

$$r = \frac{595.00 + 282.92 - 234.92}{2\sqrt{595 \times 282.92}} \quad (54)$$

$$M_Y = 30.4$$

$$= .78$$

need for taking account of the corrections, c_x and c_y , necessarily complicates the formula somewhat, but the terms are not new and the calculation is straightforward. The formula is

$$r = \frac{\Sigma x'^2 + \Sigma y'^2 - \Sigma (x' - y')^2 - 2Nc_x c_y}{2\sqrt{[\Sigma x'^2 - Nc_x^2][\Sigma y'^2 - Nc_y^2]}} \quad (55)$$

(coefficient of correlation by difference-formula,
deviations from assumed means)

Still another form of the difference-formula is often useful, especially in machine calculation. This makes use of raw or obtained scores —

$$r = \frac{N[\Sigma X^2 + \Sigma Y^2 - \Sigma (X - Y)^2] - 2(\Sigma X)(\Sigma Y)}{2\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \quad (56)$$

(coefficient of correlation by difference-formula,
calculation from raw or obtained scores)

in which $\Sigma (X - Y)^2$ is the sum of the squared differences between the two sets of scores.

TABLE 41

PROBABLE ERRORS OF THE COEFFICIENT OF CORRELATION FOR VARIOUS NUMBERS OF MEASURES (N) AND FOR VARIOUS VALUES OF r

Number of Measures	Correlation Coefficient r						
	0 00	0 10	0 20	0 30	0 40	0 50	0 60
20	1508	1493	1448	1373	1267	1131	0965
30	1231	1219	1182	1121	1035	0924	0788
40	1067	1056	1024	0971	0896	0800	0683
50	0954	0944	0915	0868	0801	0715	0610
70	0806	0798	0774	0734	0677	0605	0516
100	0674	0668	0648	0614	0567	0506	0432
150	0551	0546	0529	0501	0463	0413	0352
200	0477	0472	0458	0434	0401	0358	0305
250	0426	0421	0409	0387	0358	0319	0272
300	0389	0386	0374	0354	0327	0292	0249
400	0337	0334	0324	0307	0283	0253	0216
500	0302	0299	0290	0274	0253	0226	0193
1000	0213	0211	0205	0194	0179	0160	0137
Number of Measures	0 65	0 70	0 75	0 80	0 85	0 90	0 95
20	0871	0769	0660	0543	0419	0287	0147
30	0711	0628	0539	0444	0342	0234	0120
40	0616	0544	0467	0384	0296	0203	0104
50	0551	0486	0417	0343	0265	0181	0093
70	0466	0411	0353	0290	0224	0153	0079
100	0391	0345	0294	0242	0187	0128	0066
150	0318	0281	0241	0198	0153	0105	0054
200	0275	0243	0209	0172	0133	0091	0047
250	0246	0218	0187	0154	0118	0081	0042
300	0225	0199	0170	0140	0108	0074	0038
400	0195	0172	0148	0122	0094	0064	0033
500	0174	0154	0132	0109	0084	0057	0029
1000	0123	0109	0093	0077	0059	0041	0021

IV. THE RELIABILITY OF THE COEFFICIENT OF CORRELATION

1. The Probable Error of a Coefficient of Correlation, PE_r

The reliability of a correlation coefficient depends upon the size of the obtained r , and upon N , the size of the sample. Given these values, PE_r may be calculated from the formula

$$PE_r = \frac{.6745(1 - r^2)}{\sqrt{N}} \quad (57)$$

(probable error of a coefficient of correlation)

or it may be read more conveniently from Table 41.

For the $r = .60$ and $N = 120$ in the height-weight problem in Figure 43, the PE_r to two decimals is .04. On the assumption that this sample is representative of its population, this PE means that the chances are even (50 in 100) that the "true" r falls within the limits of $.60 \pm .04$, or between .56 and .64; and that the chances are 99 in 100 that the "true" r falls within the limits $.60 \pm 4 \times .04$, or between .44 and .76. By the "true" r (p 198) is meant that r which we should expect to find between height and weight in that population from which our sample was drawn.

To be reasonably sure that at least some degree of correlation greater than zero is present, an obtained r should be four times its PE . Given the situation in which r is exactly four times its PE , in which, for example, $r = .16$ and $PE_r = .04$, we can only be *sure* that the true r falls within the limits $.16 \pm 4 \times .04$, or between 0 and .32. If PE_r had been .06, say, the lower limit of the range would have been $-.08$, and there would have been a possibility that the true correlation is negative. It is customary, therefore, not to consider an r as *significant* (as indicative of a true correlation greater than 0) unless it is four times its PE . To be certain of at least a small degree of correlation, a low r should be five or six times its PE .

2. The Probable Error of the Difference Between Two r 's

In Chapter VIII it was shown that the reliability of the difference between two means or two σ 's can be calculated from the formulas for σ_D and PE_D (see p. 210). In the same way, the reliability of the difference between two r 's may be estimated from the PE of their difference. When the two r 's have been obtained from independent (e.g., different) samples, the PE of the difference between them is given by the formula

$$PE_{(r_1 - r_2)} = \sqrt{PE^2_{r_1} + PE^2_{r_2}} \quad (58)$$

(probable error of the difference between two coefficients of correlation r_1 and r_2 calculated from different samples)

The use of this formula may be illustrated by the following problem. Suppose that for a group of 100 eight year old boys the correlation between I.Q. and Vocabulary is .65 with a PE of .04; and that in a group of 150 eight year old girls the correlation between the same two tests is .70 with a PE of .03 (see Table 41 for PE values). The correlation coefficient is .05 higher for girls than for boys. On the assumption that our groups are fair samples of their respective populations, is this difference sufficiently large to indicate that the "true" correlation between I.Q. and Vocabulary is higher for eight year old girls than for eight year old boys? To answer this question, we must first determine the PE of the difference between the two r 's. By formula (58), $PE_{(r_1 - r_2)} = .05$, and comparing the obtained difference of .05 with its PE_D , we find that D/PE_D is 1.00. This means (Table 35) that there are 75 chances in 100 of a difference greater than zero between the "true" correlations of I.Q. and Vocabulary for eight year old boys and eight year old girls. The obtained difference of .05, therefore, is not significant. To be completely reliable, the obtained difference should have been at least $4 \times .05$ or .20. Hence, in the present case, the obtained difference is only 25% of what it should be in order to guarantee a significant difference between the correlations obtained for boys and girls.

The formulas for PE_r and $PE_{(r_1 - r_2)}$ are subject to the same restrictions, and must be interpreted with the same caution, as other standard and probable errors (Chapter VIII, p. 242). In order to possess real value as measures of reliability, PE_r and $PE_{(r_1 - r_2)}$ should be based upon r 's calculated from representative and reasonably large samples. PE 's computed for r 's obtained from small and obviously selected samples may give an entirely false picture of the value of the obtained coefficient. This is especially true when the coefficient is large. Thus an r of .90 obtained from a sample of 20 cases is of doubtful validity in spite of the fact that its PE is small (.03).*

* Fisher, R. A., *Statistical Methods for Research Workers*, 3rd edition, 1930, Chapter 6, pp. 140-177.

When (1) the reliability of the difference between the inter-correlations obtained from several tests administered to the *same* group, or when (2) the reliability of the difference between the correlations of two tests administered successively to the *same* group is desired, formula (58) is no longer strictly correct, although it will give results which may be approximately correct. Formula (58) is strictly applicable *only* when the r 's calculated from independent samples are to be compared. Methods of handling the problems stated under (1) and (2) above will be found in more advanced texts.* The problem of comparing the correlations between tests given to different groups (i.e., the problem treated in this section) is the one most often encountered.

3. Averaging Correlation Coefficients

It is a common practice to average the correlation coefficients obtained from tests given to several comparable groups, or from tests administered successively to the same group, in order to obtain a generalized picture of the relationship between the two given sets of measures. Unless precautions are taken, however, the averaging of r 's is a dubious — when not an incorrect — procedure. This is because (1) r 's do not vary along a linear scale, so that the increase in correlation from .40 to .50 does not mean the same increase in degree of relationship as an increase in correlation from .80 to .90. In fact, a correlation of .99 differs from a correlation of .95 by as much as a correlation of .71 differs from one of .01.† (2) Again, r 's may be plus and minus and hence tend to cancel each other out when averaged. Thus the average of an r of .60 and an r of $-.60$ is .00; and two substantial measures of relationship combine to give a result which indicates no real relationship. When r 's do not differ greatly in size, their arithmetic mean will yield a result which is accurate enough

* Kelley, T. L., *Statistical Method*, 1923, pp. 178-179

† Garrett, H. E., and Anastasi, A., *The Tetrad-Difference Criterion and the Measurement of Mental Traits*, Annals New York Academy of Sciences, 1932, 33, pp. 262-263.

for most purposes; but this is not true if the r 's differ widely in size or differ in sign. Thus, averaging an r of .70 and an r of .60 to obtain .65 is permissible; but averaging an r of .90 and an r of .10 to obtain .50 is not.

The safest plan is not to average r 's at all. When, for various reasons, however, averaging appears to be demanded by the problem, the best procedure is the method outlined by R. A. Fisher.* A relatively simple method which gives results which closely approximate the more exact method of Fisher is to square each r , average these squares, and extract the square root of the average thus obtained. The example given below illustrates the method:

r 's between the same tests in successive samples N equals 50 in each sample		r^2
	.20	.04
	.40	.16
	.60	.36
	.80	.64
	.90	.81
Sum	<u>2.90</u>	<u>2.01</u>
Mean	.58	.40
Mean r	.58	.63 ($\sqrt{.40}$)

The r of .63 obtained by the "squared r " method is much closer to the result to be expected from correlating the two tests in the *whole sample* (obtained by adding the N 's of the five separate samples) than is the simple average r of .58. The PE of .63 should be based upon the whole sample of 250 and not upon 50, the size of each of the component samples. Thus the PE_r of .63 is .03 (Table 41).

PROBLEMS

1. The scores upon Army Alpha and a Typewriting Test (given opposite) were made by 100 students in a typewriting class. The typewriting scores are in number of words written per minute, with

* Fisher, R. A., *Statistical Methods for Research Workers*, 3rd edition, 1930, Chapter 6, pp. 163-177.

certain penalties. Find the coefficient of correlation and PE_r . In tabulating the scores, let typing be the Y -variable, and Alpha the X -variable. Use a step-interval of 5 units for Y and of 10 units for X .

Typing (Y)	Alpha (X)	Typing (Y)	Alpha (X)	Typing (Y)	Alpha (X)
46	152	26	164	40	120
31	96	33	127	36	140
46	171	44	144	43	141
40	172	35	160	48	143
42	138	49	106	45	138
41	154	40	95	58	149
39	127	57	146	23	142
46	156	23	175	45	166
34	156	51	126	44	138
48	133	35	120	47	150
48	173	41	154	29	148
38	134	28	146	46	166
26	179	32	154	46	146
37	159	50	159	39	167
34	167	29	175	49	139
51	136	41	164	34	183
47	153	32	111	41	150
39	145	49	164	49	179
32	134	58	119	31	138
37	184	35	160	47	136
26	154	48	149	40	172
40	90	40	149	30	145
53	143	43	143	40	109
46	173	38	159	38	158
39	168	37	157	29	115
52	187	41	153	43	93
47	166	51	149	55	163
31	172	40	163	37	147
33	189	35	175	52	169
22	147	31	133	38	75
46	150	23	178	39	152
44	150	37	168	32	159
37	143	46	156	42	150
31	133				

2. In the correlation table given below, compute the coefficient of correlation and PE_r .

BOYS: AGES 4.5 TO 5.5 YEARS

Weight in Pounds (X)

Height in Inches (Y)		24-28	29-33	34-38	39-43	44-48	49-53	Totals
	45-47			1		2		3
	42-44			4	35	21	5	65
	39-41		5	87	90	7	1	190
	36-38	1	18	72	8			99
	33-35	5	15	5				25
	30-32	2						2
	Totals	8	38	169	133	30	6	384

3. In the following correlation table, compute the coefficient of correlation and the PE_r .

Army Alpha I.Q.'s

School Marks	84 and lower	85-89	90-94	95-99	100-104	105-109	110-114	115-119	120-124	125 over	Totals
90 and over				3	3	15	12	9	9	5	56
85-89				8	17	15	24	13	6	6	89
80-84			4	6	22	21	20	10	5	1	89
75-79			7	25	33	23	10	7	4		109
70-74		4	10	18	14	22	12	1	1		82
65-69	1	3	3	12	7	8	8	1			43
60-64			2	5	3	1	1				12
Totals	1	7	26	77	99	105	87	41	25	12	480

4. Compute the coefficient of correlation and PE_r between the Algebra Test scores and I.Q.'s shown in the table below.

ALGEBRA TEST SCORES

I.Q.'s		30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	Totals
	130-139				1		1		1	3
	120-129			1		1	2	1		5
	110-119	1	2	5	6	11	6	3	2	36
	100-109	3	7	9	17	13	5	1	1	56
	90-99	4	10	16	12	5	1			48
	80-89	4	9	8	2	2				25
	Totals	12	28	39	38	32	15	5	4	173

5. Compute the correlation between the two sets of scores given below
- (a) when deviations are taken from the means of the two series (use formula 51);
- (b) when the means are taken at zero. First reduce the scores by subtracting 150 from each of the scores in Test 1, and 40 from each of the scores in Test 2.

Test 1	Test 2	Test 1	Test 2
150	60	139	41
126	40	155	43
135	45	147	37
176	50	162	58
138	56	156	48
142	43	146	39
151	57	133	31
163	38	168	46
137	41	153	52
178	55	150	57

6. The following correlation coefficients represent the relationships found between two tests in five groups of approximately the same size. .65, .36, .24; .42, .27.
- Calculate the arithmetic mean of these r 's.
 - Calculate the mean by the method given on page 284.
 - Compute the mean of the following five r 's by the two methods used in (a) and (b) above: .07; .32, .41; .12; .00. Comment upon your result.
7. Find the correlation between the two sets of memory span scores given below (the first series is arranged in order of size) (a) when deviations are taken from assumed means (formula 49) (b) by the difference-method given on page 278.

Test 1 (digit span)	Test 2 (letter span)
15	12
14	14
13	10
12	8
11	12
11	9
11	12
10	8
10	10
10	9
9	8
9	7
8	7
7	8
7	6

ANSWERS

- $r = -.05$; $PE_r = .07$
- $r = .71$; $PE_r = .02$
- $r = .46$; $PE_r = .02$
- $r = .52$; $PE_r = .04$
- $r = .41$
- (a) .39; (b) .41; (c) .18 and .24
- $r = .78$

CHAPTER X

REGRESSION AND PREDICTION

I. THE REGRESSION EQUATIONS

1. The Problem of Predicting One Variable from Another

SUPPOSE that, in our group of 120 college students (p. 259), we wish to estimate a certain man's height, knowing his weight to be 150 pounds. The best possible "guess" that we can make of this man's height is the mean height of all of the men who fall in the 150-159 weight-interval. In Figure 43 the mean height of the nine men in this column is 68.9 inches, which is, therefore, the most likely height of a man who weighs 150 pounds. In the same way, the most probable height of a man who weighs 165 pounds is 68.9 inches, the mean height of the five men who fall in weight-column 160-169 pounds. And, in general, the most probable height of any man in the group is the *mean* of the heights of *all* of the men who weigh the same (or approximately the same) as he, i.e., who fall within the same weight-column.

Shifting to weight, we can make the same kind of estimates. Thus, the best possible "guess" that we can make of a man's weight knowing his height to be 67 inches is 135.1 pounds — viz, the mean weight of the 33 men who fall in the height-interval 66-67 inches. Again, in general, the most probable weight of any man in the group is the *mean* weight of *all* of the men who are of the same (or approximately the same) height.

Our illustration shows that from the scatter diagram alone, it is possible to "predict" one variable from another. But the prediction is rough, and is obviously subject to a large "error of estimate." * Moreover, while we have made use of the fact

* See page 300.

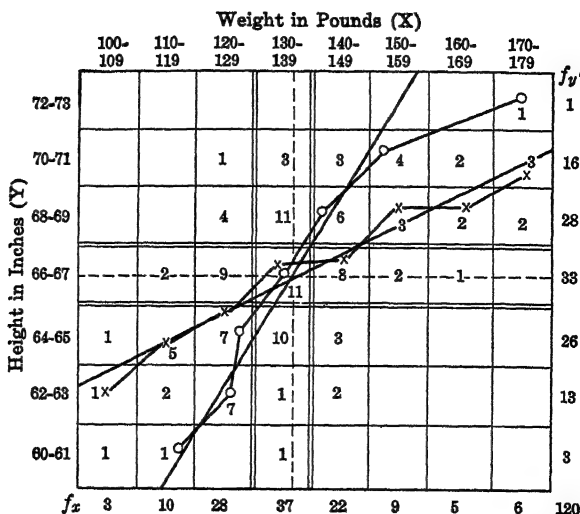


FIG. 47. Illustrating Positions of Regression Lines and Calculation of the Regression Equations (see Fig. 46, p. 267).

$$r = .60 \pm .04 \text{ (Fig. 46)}$$

$$M_X = 136.3 \text{ lbs.}$$

$$M_Y = 66.5 \text{ inches}$$

For plotting on the chart, regression equations are written with σ_x and σ_y in step-units, viz. —

$$\left. \begin{aligned} y &= .51x \\ x &= .71y \end{aligned} \right\} \text{ see p. 296.}$$

Calculation of Regression Equations

I. Deviation Form

$$(1) \quad \bar{y} = .60 \times \frac{2.62}{15.54} x = .10x \quad (59)$$

$$(2) \quad \bar{x} = .60 \times \frac{15.54}{2.62} y = 3.56y \quad (60)$$

II. Score Form

$$(1) \quad Y - 66.5 = .10(X - 136.3) \text{ or } \bar{Y} = .10X + 52.9 \quad (61)$$

$$(2) \quad X - 136.3 = 3.56(Y - 66.5) \text{ or } \bar{X} = 3.56Y - 100.4 \quad (62)$$

Calculation of Standard Errors of Estimate

$$\sigma_{(\text{est } Y)} \text{ or } \sigma_{Y.X} = 2.62 \sqrt{1 - .60^2} = 2.10 \text{ inches} \quad (63)$$

$$\sigma_{(\text{est } X)} \text{ or } \sigma_{X.Y} = 15.54 \sqrt{1 - .60^2} = 12.43 \text{ pounds} \quad (64)$$

that the means are the most probable points in our arrays (columns or rows), we have made no use of our knowledge concerning the correlation between the two variables. It seems fairly clear from Figure 47 that the two "regression"* lines drawn in on the chart describe more regularly, and in a more generalized fashion than do the series of short straight lines joining the means, the relationship between height and weight over the whole range (see also p. 264). Furthermore, the regression lines are definitely determined by the correlation between height and weight and their degree of separation indicates the size of the correlation coefficient (p. 262). A knowledge of the equations of these lines, therefore, would certainly seem to be desirable. For then, given the weight (X) of any man in our group, we should be able by substituting in the equation connecting Y and X to find directly this man's most probable height. The task of the next section will be to develop equations for these regression lines by means of which precise predictions from X to Y or from Y to X can be effected.

2. The Two Regression Equations in Deviation Form

The equations of the two regression lines in a correlation table (which were first derived by Karl Pearson) describe the straight lines which "best fit" the means of the successive columns and rows in a correlation table. Using as his criterion of "best fit" the method of least squares,† Pearson worked out the equation of the line which goes through, or as close as possible to, more of the column-means than any other straight line;

* The term "regression" was first used by Francis Galton with reference to the inheritance of stature. Galton found that children of tall parents tend to be less tall, and children of short parents less short, than their parents. In other words, the heights of the offspring tend to "move back" toward the mean height of the general population. This tendency toward maintaining the "mean height" Galton called the principle of regression, and the line describing the relationship of height in parent and offspring was called a "regression line." The term is still employed, although its original meaning of "stepping back" to some stationary average is no longer implied.

† For an elementary mathematical treatment of the application of least squares to the problem of fitting regression lines, see Holzinger, K. J., *Statistical Methods for Students of Education*, 1928, pp. 159 ff.

and also the equation of the line which goes through, or as close as possible to, more of the row-means than any other straight line. In a mathematical sense, therefore, these two lines are "best fitting," the one to the observations of the columns and the other to the observations of the rows.

The equation of the first regression line, the line drawn through the crosses in Figure 47, is as follows:*

$$\bar{y} = r \frac{\sigma_y}{\sigma_x} \cdot x \quad (59)$$

*(regression equation of y on x, deviations taken from
the means of Y and X)*

The factor $r \frac{\sigma_y}{\sigma_x}$ is called the *regression coefficient*, and is often replaced in (59) by the term b_{yx} or b_{12} so that formula (59) may be written $\bar{y} = b_{yx} \cdot x$, or $\bar{y} = b_{12} \cdot x$.

If we substitute in formula (59) the values of r , σ_y , and σ_x , obtained from Figure 46, we have

$$\bar{y} = .60 \times \frac{2.62}{15.54} x, \text{ or } \bar{y} = .10x$$

This equation gives the relationship of deviations in height to deviations in weight. When $x = 1.00$, $\bar{y} = .10$; hence a deviation of one pound from the mean of the X 's (weight) is accompanied by a deviation of .10 inch from the mean of the Y 's (height). The man who stands one pound *above* the mean weight of the group, therefore, is most probably .10 inch *above* the mean height. Since his weight is 137.3 pounds ($136.3 + 1.00$), his height is most probably 66.6 inches ($66.5 + .10$). Again, the man who weighs 120 pounds, i.e., is 16.3 pounds *below* the mean of the group, is most probably 65 inches tall — about 1.6 inches *below* the mean height of the group. To get this last value, substitute $x = -16.3$ in the equation above to get $\bar{y} = -1.63$, and refer this value to its mean. In general, this regression equation tells us that the most probable deviation of an individual in our group from the $M(ht)$ is just .10 of his deviation from the $M(wt)$.

* The bar over the y means that our estimate is an average value.

The regression equation $\bar{y} = r \frac{\sigma_y}{\sigma_x} \cdot x$ gives the relationship between y and x in *deviation form*. This designation is necessary because the two variables are expressed as deviations from their respective means (i.e., as x and y); hence, for a given *deviation* from M_x the equation gives the most probable accompanying *deviation* from M_y .

The equation of the second regression line, the line drawn through the means of the rows in Figure 47, is written

$$\bar{x} = r \frac{\sigma_x}{\sigma_y} \cdot y \quad (60)$$

(regression equation of x on y , deviations taken from the means of X and Y)

As in the first regression equation, the regression coefficient

$r \frac{\sigma_x}{\sigma_y}$ is often replaced by the expression b_{xy} or b_{21} and formula (60) written $\bar{x} = b_{xy} \cdot y$ or $\bar{x} = b_{21} \cdot y$.

If we substitute for r , σ_x , and σ_y , in formula (60), we have

$$\bar{x} = .60 \times \frac{15.54}{2.62} y \text{ or } \bar{x} = 3.56y$$

from which it is evident that a deviation of 1 inch from the $M(ht)$, or from 66.5 inches, is accompanied by a deviation of 3.56 pounds from the $M(wt)$, or from 136.3 pounds. Expressed generally, the most probable deviation of *any man* from the mean weight is just 3.56 times as great as his deviation from the mean height. Accordingly, a man 67 inches tall or .5 inch *above* the mean height ($66.5 + .5 = 67$) most probably weighs 138.1 pounds, or is 1.8 pounds *above* the mean weight ($136.3 + 1.8$). (Substitute $y = .5$ in the equation and $\bar{x} = 1.8$.)

Equation $\bar{x} = r \frac{\sigma_x}{\sigma_y} \cdot y$ gives the relationship between x and y in *deviation form*. More precisely, it gives the most probable *deviation* of an X -measure from M_x corresponding to a known *deviation* in the Y -measure from M_y .

Although both of the regression equations given above involve x and y , the two equations cannot be used interchangeably — neither can be used to predict *both* x and y . This is an important fact which the student must understand clearly and constantly bear in mind. The first regression equation $\bar{y} = r \frac{\sigma_y}{\sigma_x} \cdot x$ can be used *only* when y is to be predicted from a given x (when y is the “dependent” variable)*; while the second equation $\bar{x} = r \frac{\sigma_x}{\sigma_y} \cdot y$ can be used *only* when x is to be predicted from a known y (when x is the “dependent” variable). There are always two regression equations in a correlation table, the one through the means of the columns and the other through the means of the rows, unless the correlation is 1.00 or -1.00.

When $r = 1.00$, $\bar{y} = r \frac{\sigma_y}{\sigma_x} \cdot x$ becomes $\bar{y} = \frac{\sigma_y}{\sigma_x} \cdot x$ or $\bar{y}\sigma_x = x\sigma_y$.

Also, when $r = 1.00$, $\bar{x} = r \frac{\sigma_x}{\sigma_y} \cdot y$ becomes $\bar{x} = \frac{\sigma_x}{\sigma_y} \cdot y$ or $\bar{x}\sigma_y = y\sigma_x$. In short, when the correlation is perfect (± 1.00), the two equations are identical and the two regression lines coincide. To illustrate this situation, suppose that the correlation between height and weight in Figure 47 were perfect. The

first regression equation would then be $\bar{y} = 1.00 \times \frac{2.62}{15.54} x$ or

$\bar{y} = .17x$, and the second, $\bar{x} = 1.00 \times \frac{15.54}{2.62} y$, or $\bar{x} = 5.93y$.

Algebraically, the equation $x = 5.93y$ is equal to $y = .17x$; for if we put $x = \frac{y}{.17}$, $x = 5.93y$. When $r = \pm 1.00$ there is only

one equation and a *single* regression line. Moreover, if $r = 1.00$, and in addition $\sigma_x = \sigma_y$, the single regression line makes an angle of 45° with the horizontal axis, since $y = x$.

* The dependent variable takes its value from the other (independent) variable in the equation. For example, in the equation $y = 3x$, y “depends” for its value upon x ; and hence y is the dependent variable.

3. Plotting the Regression Lines in a Correlation Table *

In Figure 47, the coördinate axes have been drawn in on the correlation table through the means of the X - and Y -distribu-

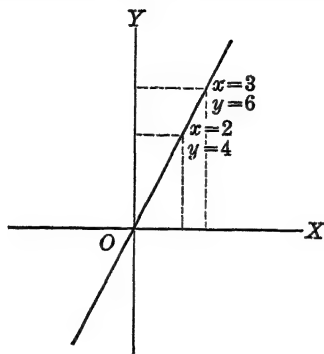


FIG. 48. Plot of the straight line, $y = 2x$.

* A brief review of the equation of a straight line, and of the method of plotting a simple linear equation is given here in order to simplify the plotting of the regression equations.

In Figure 48, let X and Y be coordinate axes, or axes of reference. Now suppose that we are given the equation $y = 2x$ and are required to represent the relation between x and y graphically. To do this we assign values to x in the equation and compute the corresponding values of y . When $x = 2$, for example, $y = 2 \times 2$ or 4; when $x = 3$, $y = 2 \times 3$ or 6. In the same way, given any x -value we can compute the value of y which will "satisfy" the equation, that is, make the left side equal to the right. If the series of x and y values found from the equation are plotted on the diagram with respect to the X - and Y -coördinates (as in Fig. 48) they will be found to fall along a straight line. This straight line pictures the relation $y = 2x$. It goes through the origin, since when $x = 0$, $y = 0$. The equation $y = 2x$ represents, then, a straight line which passes through the origin; and the relation of its coordinates (points lying along the line) is such that $\frac{y}{x}$, called the *slope* of the line, is always equal to 2.

The general equation of any straight line which passes through the origin may be written $y = mx$, where m is the slope of the line. If we replace m in the general formula by $r \frac{\sigma_y}{\sigma_x}$ it is clear that the regression line in *deviation form*, namely, $y = r \frac{\sigma_y}{\sigma_x} x$, is simply the equation of a straight line which goes through the origin. For the same reason, when the general equation of a straight line through the origin is written $x = my$, $x = r \frac{\sigma_x}{\sigma_y} y$ is also seen to be a straight line through the origin, its slope being $r \frac{\sigma_x}{\sigma_y}$.

tions. The vertical axis is drawn through 136.3 pounds (M_{wt}), and the horizontal axis through 66.5 inches (M_{ht}). These axes intersect close to the center of the chart. Equations (59) and (60) define straight lines which pass through the origin or point of intersection of these coordinate axes. For this reason, it is a comparatively simple task to plot in our regression lines on the correlation chart with reference to the given coordinate axes.

Correlation charts are usually laid out with equal distances representing the X and Y step-intervals (the printed correlation charts are always so constructed) although the step-intervals expressed in terms of the variables themselves may be, and often are, unequal. This is true in Figure 47. In this diagram, the step-intervals in X and Y are drawn equal, although the actual step-interval for height is 2 inches, and the actual step-interval for weight is 10 pounds. Because of this difference in step-length in the two variables it is very important that we express σ_x and σ_y in our regression equations in *step-intervals*, before plotting the regression lines in on the chart. Otherwise we must equate our X and Y step-intervals by laying out our diagram in such a way as to make the X -step five times the Y -step. This latter method of equating step-intervals is impractical, and is rarely used, since all we need do in order to use correlation charts containing equal step-intervals is to drop out the step-lengths in formulas (59) and (60). When this is done, and the step-interval, *not* the score, is the unit, the first regression equation becomes

$$y = .60 \frac{1.31}{1.55} x \text{ or } y = .51x$$

and the second

$$x = .60 \frac{1.55}{1.31} y \text{ or } x = .71y$$

Since each regression line goes through the origin, only one other point (besides the origin) is needed in order to determine it. In the first regression equation, if $x = 10$, $y = 5.1$; and the two points (0, 0) and (10, 5.1) locate the line. In the second regression equation, if $y = 10$, $x = 7.1$; and the two points (0, 0) and (7.1, 10) determine the second line. In plotting points

on a diagram any convenient scale may be employed. A millimeter rule is useful.

It is important for the student to remember that when the two σ 's are expressed in step-units, regression equations do *not* give the relationship between the X and Y score deviations. These special forms of the regression equations should not be used except when plotting the equations on a correlation chart. Whenever the most probable deviation in the one variable corresponding to a known deviation in the other is wanted, formulas (59) and (60), in which the σ 's are expressed in *score units*, must be employed.

4. The Regression Equations in Score Form

In the last sections it was pointed out that formulas (59) and (60) give the equations of the regression lines in deviation form — that values of x and y substituted in these equations are deviations from the means of the X and Y distributions, and not actual scores. While the equations in *deviation form* are actually all that one needs in order to predict from one variable to another, it is decidedly convenient to be able to estimate an individual's *actual score* in Y , say, directly from his score in X without first converting the X -score into a deviation from M_X . This can be done by using the *score form* of the regression equations. The conversion of deviation form to score form is made as follows. Denoting the mean of the Y 's by M_Y and any Y -score simply by Y , we may write the deviation of any individual from the mean as $Y - M_Y$ or, in general, $y = Y - M_Y$. In the same way, $x = X - M_X$ when x is the deviation of any X -score from the mean X . If we substitute $Y - M_Y$ for y , and $X - M_X$ for x , in formulas (59) and (60), the two regression equations become

$$Y - M_Y = r \frac{\sigma_y}{\sigma_x} (X - M_X) \text{ or } \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - M_X) + M_Y \quad (61)$$

and

$$X - M_X = r \frac{\sigma_x}{\sigma_y} (Y - M_Y) \text{ or } \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - M_Y) + M_X \quad (62)$$

(regression equations of Y on X and X on Y in score form)

The two equations are now said to be in *score form*, since the X and Y in both equations represent *actual scores*, and *not deviations from the means* of the two distributions.

If we substitute in (61) the values of M_Y , r , σ_y , σ_x and M_X obtained from Figure 47, the regression of height on weight in score form becomes

$$\bar{Y} = .60 \times \frac{2.62}{15.54} (X - 136.3) + 66.5$$

or clearing of fractions

$$\bar{Y} = .10X + 52.9$$

To illustrate the use of this equation, suppose that a man in our group weighs 160 pounds and we wish to estimate his most probable height. Substituting 160 for X in the equation, $\bar{Y} = 69$ inches; and accordingly, the most probable height of a man who weighs 160 pounds is 69 inches.

If the problem is to predict weight instead of height, we must use the second regression equation, formula (62). Substituting for M_X , r , σ_x , σ_y , and M_Y in (62) we have

$$\bar{X} = .60 \times \frac{15.54}{2.62} (Y - 66.5) + 136.3$$

or

$$\bar{X} = 3.56Y - 100.4$$

Now if a man is 71 inches tall, we find, on putting 71 in the equation for Y , that $\bar{X} = 152.4$. Hence the most probable weight of a man who is 71 inches tall is about 152 pounds.

It may seem strange to the student, perhaps, that we should talk of "predicting" a man's height from his weight, when the heights and weights of all of the 120 men in our group are already known. When we have measures of both height and weight it is, ordinarily, unnecessary to estimate one from the other. But suppose that all we know about a given individual is his weight and the fact that he falls within the age-range of our group of 120 men. Since we know the correlation between height and weight to be .60, it is possible from the regression equation to predict the most probable height of our subject in

lieu of actually measuring him. Furthermore, the regression equation may be employed to estimate the height of any man in the population from which our group is chosen, provided our sample is a random selection of the larger group. A regression equation holds, of course, only for the population from which the sample group was drawn. We cannot estimate the heights of children or of women from a regression equation which describes the relationship between height and weight in men between the ages of 18 and 25 years (the age-range of the students in our group). Conversely, we cannot expect a regression equation established for elementary school children to hold for older groups.

Height and weight, since they are both easily measured, do not show so clearly, perhaps, the value of the regression equation as do other and more complex traits. These characteristics were chosen for our "model" problem because they are objective attributes, the meaning of which is definite. Let us now consider a problem of more direct psychological interest. Suppose that in a group of 300 high school children of nearly the same age, the correlation between group test M.A. obtained at the beginning of school and average grade made in the first year of high school is .60. Now if we know the group test M.A. of a child who enters school the next year, it is possible to estimate his probable scholastic performance by means of the regression equation between M.A. and grades obtained from the previous year's class. Estimates of this sort have proved useful in educational prognosis and guidance.* The same is true of vocational guidance; we are often able, on the basis of regression equations already established, to predict from a test battery the probable success of an individual who contemplates entering a certain trade or profession †. Advice on such a basis is certainly better than snap judgment.

* Pintner, R., *Intelligence Testing, Methods and Results*, 1931, Chapters 10, 11, 12.

Edgerton, H. A., *Academic Prognosis in the University*, Educational Psychology Monographs, 27, 1930

† Hull, C. L., *Aptitude Testing*, 1928, Chapter 5, pp. 179-183.

II. THE RELIABILITY OF PREDICTIONS MADE FROM REGRESSION EQUATIONS

1. The Standard Error of Estimate

The values of X and Y "predicted" from regression equations have been constantly referred to as being the "most probable" values of the one variable accompanying the given value of the other. In order to show just how probable such estimates are it is necessary that we calculate their standard errors of estimate. The accuracy with which we are able to predict Y -scores from equation (61) is given by the formula

$$\sigma_{(\text{est } Y)} = \sigma_y \sqrt{1 - r^2} \quad (63)$$

(standard error of a Y -score predicted from equation (61))

in which σ_y is the σ of the Y distribution, and r is the coefficient of correlation. The subscript "est." is used to distinguish this σ from the σ of the distribution, the σ of the mean, etc.

From formula (61) we have found that the most probable height of a man weighing 160 pounds is 69 inches. The reliability of this prediction is obtained by substituting $\sigma_{(M)}$ and r in formula (63) to find

$$\sigma_{(\text{est } Y)} = 2.62 \sqrt{1 - .60^2} = 2.1 \text{ inches}$$

We may now say that the most probable height of a man weighing 160 pounds is 69 inches with a $\sigma_{(\text{est } Y)}$ of 2.1 inches; or that the chances are 68 in 100 that the actual height of the given individual falls within the limits 69 ± 2.1 or between about 67 inches and 71 inches. We may be practically certain that the height of this man will fall within the limits $69 \pm 3 \times 2.1$ or between 63 inches and 75 inches (approximately).

The degree of accuracy with which X -scores can be predicted from (62) is given by the formula

$$\sigma_{(\text{est } X)} = \sigma_x \sqrt{1 - r^2} \quad (64)$$

(standard error of an X -score predicted from equation (62))

in which σ_x is the σ of the X distribution, and r is the correlation.

It was found on page 298 that the most probable weight of a man in our group who is 71 inches tall is 152.4 pounds. The $\sigma_{(\text{est})}$ of this prediction from (64) is

$$\sigma_{(\text{est } X)} = 15.54 \sqrt{1 - .60^2} = 12.4 \text{ pounds}$$

Hence, the most probable weight of any man 71 inches tall, in our group or in the population from which it is drawn, is 152.4 pounds, with a $\sigma_{(\text{est})}$ of 12.4 pounds. The chances, therefore, are 68 in 100 that the measured weight of this man falls within the limits 152.4 ± 12.4 or between 140 pounds and 165 pounds. We can only be certain that his weight falls within the limits $152.4 \pm 3 \times 12.4$ or between 115 pounds and 190 pounds (to the nearest one pound).

2. The Probable Error of Estimate

Instead of $\sigma_{(\text{est})}$ the probable error of estimate may be employed for estimating the accuracy of a prediction. $PE_{(\text{est})}$ is obtained from $\sigma_{(\text{est})}$ by multiplying $\sigma_{(\text{est})}$ by the constant .6745 (see p. 114). Thus

$$PE_{(\text{est } Y)} = .6745 \sigma_y \sqrt{1 - r^2} \quad (65)$$

and

$$PE_{(\text{est } X)} = .6745 \sigma_x \sqrt{1 - r^2} \quad (66)$$

(probable errors of Y- and X-scores predicted from regression equations)

The height of a man who weighs 160 pounds has been estimated to be 69 inches with a $\sigma_{(\text{est})}$ of 2.1 inches. The $PE_{(\text{est})}$ of this predicted height is $.6745 \times 2.1$ or 1.4 inches. The chances are even, therefore, or 50 in 100 that the actual height of this man falls within the limits 69 ± 1.4 inches or between 67.6 inches and 70.4 inches.

$PE_{(\text{est } X)}$ is used in the same way as $PE_{(\text{est } Y)}$. We found that the estimated weight of a man 71 inches tall was 152.4 pounds with a $\sigma_{(\text{est})}$ of 12.4 pounds. Hence, the $PE_{(\text{est } X)}$ of

this weight is $.6745 \times 12.4$ or 8.4 pounds, and the chances are even that this man's actual weight lies within the limits 152.4 ± 8.4 or between 144 pounds and 161 pounds (to the nearest one pound).

The formulas for $\sigma_{(est)}$ and $PE_{(est)}$ measure the error made in taking predicted, instead of actual, X and Y measures. If $r = 1.00$, $\sqrt{1 - r^2}$ is 0, and both $\sigma_{(est)}$ and $PE_{(est)}$ are zero. There is, in consequence, *no* error of estimate and each person's measurement is predicted exactly. On the other hand, when $r = .00$, $\sqrt{1 - r^2} = 1.00$, and the error of estimate is equal to the σ of the distribution. When this happens, the regression equation is of no value as an aid in enabling us the better to predict scores, since each person's most probable score is simply the mean. When $r = .00$ all that we can say definitely is that a subject's score lies *somewhere* in the distribution of Y 's or X 's. But just where we cannot tell.

It is clear from formulas (63) and (64) that the accuracy of a prediction from a regression equation depends directly upon the σ 's of the two distributions (σ_y or σ_x) and upon the degree of correlation between the two sets of measures. If the variability (σ_y) of Y is small, and the correlation between Y and X high (e.g., .90), values of Y can be predicted from known values of X with a comparatively high degree of accuracy. However, when the variability of a test is large, or the correlation low (or when both conditions obtain), prediction from regression equations becomes so unreliable as to be almost valueless. Even when the correlation is fairly high, predictions will often have an uncomfortably large error of estimate. Thus we have seen that in spite of the $r = .60$ between height and weight (Fig. 46), we can only predict a man's weight, knowing his height, with a $PE_{(est\ x)}$ of about 8 pounds (p. 302). In other words, our predicted weights will in 50% of the cases be in error by as much as 8 pounds. Prediction of height from weight is considerably better than the prediction of weight from height because of the relatively small "spread" (σ) of the measurements of height. Predicted heights, for example, will, in 50% of the

cases, be in error by not more than 1.4 inches. In the correlation problem given later in Figure 49, the estimates of I.Q. are quite accurate. The high correlation offsets the fact that the variability is fairly large.

When an investigator uses the regression equations for purposes of prediction, he should always give the $\sigma_{(\text{est})}$, or the $PE_{(\text{est})}$ of his estimated scores. The value of a prediction depends, first of all, upon the size of the error of estimate; but it also depends upon the fineness of the units of measurement, and upon the purposes for which the prediction is made.

III. THE EFFECT UPON THE CORRELATION COEFFICIENT OF THE RANGE OF TALENT IN THE GROUP

Suppose that the correlation between two tests in a small group of 50 sixth grade children has been found to be .50. How will the correlation between the same tests in a larger group, e.g., a group of 200 children in the sixth grade or 200 children spread over grades six, seven, and eight, compare with the obtained r of .50? More generally, knowing the correlation between two tests in a small (and homogeneous) group, can we estimate or predict the probable correlation in a large (and heterogeneous) group?

The problem of the effect upon r of the range of talent (size of σ_x and σ_y) within the group being studied often arises in correlational work. It becomes important, for example, when one wishes to go beyond the correlation obtained in the sample with which he is working and generalize (estimate the r) for a larger — and more heterogeneous — group; or when r 's between the same tests obtained in different ranges are to be compared. A formula for estimating the correlation between two tests in a heterogeneous group when we know the correlation between the tests in a homogeneous group may be developed in the following way. Let $\sigma_{(\text{est } Y_s)}$ be the standard error of estimate in the small or homogeneous group; and $\sigma_{(\text{est } Y_L)}$ be the standard error of estimate in the large or hetero-

geneous group (Y is the *dependent* variable, p. 294). Then, on the assumption that our tests are as effective in the *wide* as in the *narrow* range, $\sigma_{(est\ Y_s)} = \sigma_{(est\ Y_l)}$, or, by formula (63), page 300

$$\sigma_{y_s} \sqrt{1 - r_{x_s y_s}^2} = \sigma_{y_l} \sqrt{1 - r_{x_l y_l}^2}$$

and

$$\frac{\sigma_{y_s}}{\sigma_{y_l}} = \frac{\sqrt{1 - r_{x_l y_l}^2}}{\sqrt{1 - r_{x_s y_s}^2}} \quad (67)$$

(formula for estimating correlation in a wide range from a knowledge of the correlation in a narrow range)

in which σ_{y_s} is the standard deviation of Y in the small group, or in the narrow range; σ_{y_l} is the standard deviation of Y in the large group, or in the wide range; $r_{x_s y_s}$ = the correlation in the small group, and $r_{x_l y_l}$ = the correlation in the large group.

To illustrate formula (67), suppose that in a small group $\sigma_{y_s} = 10$ and $r_{x_s y_s}$ is .50. What would the r between the same two tests probably be in a group in which $\sigma_{y_l} = 15$ — in which σ_{y_l} is 50% larger than σ_{y_s} ? Substituting $\sigma_{y_s} = 10$, $\sigma_{y_l} = 15$, and $r_{x_s y_s} = .50$ in (67), we have

$$\frac{10}{15} = \frac{\sqrt{1 - r_{x_l y_l}^2}}{\sqrt{1 - .25}}$$

Squaring both sides of this equation, and solving, we have $r_{x_l y_l} = .82$. Hence, the r of .50 in the narrow range is equivalent to an r of .82 in the wide range. For the r 's between two tests to be strictly comparable, therefore, the variability (σ) within the groups from which they were computed must be the same — or approximately the same.

If X and not Y is the dependent variable, formula (67) becomes

$$\frac{\sigma_{x_s}}{\sigma_{x_l}} = \frac{\sqrt{1 - r_{x_l y_l}^2}}{\sqrt{1 - r_{x_s y_s}^2}} \quad (68)$$

(formula for estimating correlation in a wide range from a knowledge of the correlation in a narrow range)

Both formulas (67) and (68) are open to the objection that each takes account of only *one* standard deviation (σ_y or σ_x) in estimating the probable increase in r with increase in range of talent. If the increase in σ_y , as the group becomes more heterogeneous, however, is accompanied by a proportionate increase in σ_x (or *vice versa*) formulas (67) and (68) should hold reasonably well. Experimental trial of these formulas has given results closely in accord with theoretical expectation.*

IV. THE COMPLETE SOLUTION OF A CORRELATION PROBLEM

The complete solution of a second correlation problem will be found in Figure 49. The purpose of another "model" problem is to strengthen the student's grasp on correlational techniques by having him work through the process of finding r and of calculating the regression equations upon a new set of data. One often fails to understand fully certain points in the solution of a given problem; and the solution of another entirely different problem may succeed in clearing up these difficulties.

The problem in Figure 49 is to find the relationship between the I.Q.'s of 136 children (of the same chronological age) determined from two individual intelligence examinations. The correlation table has been constructed from a scattergram as described on page 258. The test given first is the X -variable, and the test given second the Y -variable. The calculation of the two means, and of c_x , c_y , σ_x , and σ_y , covers familiar ground, is given in detail on the chart, and need not be repeated.

The product-deviations in the $\Sigma x'y'$ column have been taken from column 115-119 (the column in which the AM_x has been taken) and from row 115-119 (the row in which the AM_y has been taken). The entries in the $\Sigma x'y'$ column have been calculated by the shorter method described on page 269; that is, each cell entry in a given row was multiplied first by its

* Peters, C C, and VanVoorhis, W R, *Statistical Procedures and their Mathematical Bases*, 1935, pp. 178-180.

See also Kelley, T. L., *Statistical Method*, 1923, pp. 224-225.

x -deviation (x') and the sum of these partial deviations entered in the column $\Sigma x'$. The $\Sigma x'$ entries were then "weighted," once for all, by the y' of the whole row. To illustrate, in the first row reading from left to right, $(1 \times 5) + (1 \times 6) + (1 \times 7)$

I. Q. First Test (X)

90-94 95-99 100-104 105-109 110-114 115-119 120-124 125-129 130-134 135-139 140-144 145-149 150-154

I. Q. Second Test (Y)

	f_y	y'	$f y'$	$f y'^2$	$\Sigma x'$	$\Sigma x' y'$
155-159	1	1	1	1	3	8
150-154	1	1	1	1	2	7
145-149	1	1	1	1	2	6
140-144	4	3	1	1	8	5
135-139	1	4	1	1	6	4
130-134	1	1	3	1	7	3
125-129	1	7	1	1	13	2
120-124	3	2	3	1	13	1
115-119	3	16	5	2	26	0
110-114	2	7	5	2	19	-1
105-109	2	3	3	2	12	-2
100-104	4	5	3	1	15	-3
95-99	2	2	1	1	6	-4
90-94	2	1			3	-5
85-89	1				1	-6

f_x	3	3	8	14	21	26	14	14	8	11	7	4	3	136	
x'	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7		
$f x'$	-15	-12	-24	-28	-21	(-100)	14	28	24	44	35	24	21	190	=90
$f x'^2$	75	48	72	56	21		14	56	72	176	175	144	147		=1056
$\Sigma y'$	-16	-13	-24	-36	-21	-5	2	17	16	42	32	26	21		=41
$\Sigma x' y'$	80	52	72	72	21		2	34	48	168	160	156	147		=1012

$$c_y = \frac{41}{136} = .30$$

$$c_x = \frac{90}{136} = .66$$

$$c_y^2 = .09$$

$$c_x^2 = .44$$

$$c_{y^2} = .30 \times 5 = 1.50$$

$$c_{x^2} = .66 \times 5 = 3.3$$

$$M_Y = 117 + 15 = 118.5$$

$$M_X = 117 + 3.3 = 120.3$$

$$r = \frac{1012 - .30 \times 96}{2.95 \times 2.71} = .91 \pm .01 \text{ (Table 41)}$$

$$\sigma_y = \sqrt{\frac{1195}{136}} - .09 \times \text{step} \quad \sigma_x = \sqrt{\frac{1056}{136}} - .44 \times \text{step}$$

$$= 2.95 \times 5$$

$$= 2.71 \times 5$$

$$= 14.75$$

$$= 13.55$$

Calculation of Regression Equations

Calculation of $PE_{(est.)}$

I. Deviation Form

$$\bar{y} = .91 \times \frac{14.75}{13.55} x$$

$$= .99x$$

$$\bar{x} = .91 \times \frac{13.55}{14.75} y$$

$$= .84y$$

II. Score Form

$$Y - 118.5 = .99(X - 120.3)$$

$$\bar{Y} = .99X - 6$$

$$X - 120.3 = .84(Y - 118.5)$$

$$\bar{X} = .84Y + 20.8$$

$$PE_{(est. \ x)} = .6745 \times 14.75 \sqrt{1 - (.91)^2}$$

$$= 4.12$$

$$PE_{(est. \ y)} = .6745 \times 13.55 \sqrt{1 - (.91)^2}$$

$$= 3.79$$

Examples:

$$\text{If } X = 100$$

$$\bar{Y} = 99 - .6 = 98.4$$

$$= 98 \pm 4 \text{ (approx.)}$$

$$\text{If } X = 120$$

$$\bar{Y} = 119 - .6 = 118.4$$

$$= 118 \pm 4 \text{ (approx.)}$$

$$\text{If } Y = 130$$

$$\bar{X} = 109.2 + 20.8 = 130$$

$$= 130 \pm 4 \text{ (approx.)}$$

FIG. 49. Calculation of the Correlation Between the I.Q.'s Received by 136 Children of the Same Chronological Age in Two Individual Intelligence Examinations.

or 18 constitutes the $\Sigma x'$ entry. The x' 's are 5, 6, and 7, respectively, and may be read from the x' row at the bottom of the correlation table. Since the common y' is 8, the final $\Sigma x'y'$ entry is 144. Again, in the eighth row reading down from the top, $(3 \times -1) + (2 \times 0) + (3 \times 1) + (3 \times 2) + (1 \times 3) + (1 \times 4)$ or 13 makes up the $\Sigma x'$ entry. The y' of this row is 1, and the final $\Sigma x'y'$ entry is 13. To take still a third example, in the eleventh row reading down, $(2 \times -3) + (3 \times -2) + (3 \times -1) + (2 \times 0) + (2 \times 1)$ or -13 is the $\Sigma x'$ entry; and since the common y' is (-2) the $\Sigma x'y'$ entry is 26.

Three checks of the calculations (see p. 267) upon which r , σ_x , and σ_y are based are given in Figure 49. Note that $\Sigma fx' = \Sigma x'$; and that, when the $\Sigma x'y'$'s are recalculated, at the bottom of the chart $\Sigma fy' = \Sigma y'$ and $\Sigma x'y' = \Sigma x'y'$. When the $\Sigma x'y'$ have been checked, the calculation of r by formula (49) is a matter of substitution. Note carefully that c_x , c_y , σ_x , σ_y are all left in *units of step-interval* in the formula for r (p. 270).

The regression equations in *deviation form* are given under I, and the two lines which these equations define have been plotted on the chart. It should be noted that these equations

may be plotted as they stand, since the step-interval is the same for X and Y (p. 296). In the routine solution of a correlational problem, it is not strictly necessary to plot in the regression lines on the chart. They are often of value, however, in indicating whether the means of the X - and Y -arrays can be fairly represented by straight lines, that is, whether regression is "linear." If the relation between X and Y is not straight-line, other methods of calculating the correlation must be employed (p. 393).

Several examples have been given in Figure 49 to illustrate the use of the regression equations in *score form* in prediction. Thus, an I.Q. of 100 on the first test (X) is most probably accompanied by an I.Q. of 98 on the second test with a $PE_{(est\ Y)}$ of about four points. The chances are 50 in 100 that any child who earns an I.Q. of 100 on the first test will earn an I.Q. between 94 and 102 on the second test (98 ± 4). An I.Q. of 120 on the first test is most probably accompanied by an I.Q. of 118 ± 4 on the second test. In general, all predicted I.Q.'s on the Y -test may be considered to have the same error of estimate (four points) no matter where they fall on the scale.

While errors of estimate, $\sigma_{(est)}$ and $PE_{(est)}$, are most often employed to give the reliability of a *specific* predicted measure, they may also be interpreted in a more general fashion. Thus a $PE_{(est\ Y)}$ of four points means that one-half of the I.Q.'s in test Y failed of perfect correlation with the I.Q.'s in test X by ± 4 points or more, while the other half failed of perfect correlation by less than ± 4 points. Again, 50% of the I.Q.'s predicted on test Y from test X will be in error by four points or less; the other 50% will be in error by more than four points.

In most correlation problems we are chiefly interested in predicting the scores on only *one* test. Usually, the Y -test is the dependent, and the X -test the independent variable. For illustrative purposes, however, an example is given in Figure 49 of the prediction of an I.Q. in X from an I.Q. in Y . If the I.Q. in test Y is 130, the most probable I.Q. in test X is 130 ± 4 ;

hence, the chances are even that the actual I.Q. in X will fall between the limits 126 and 134.

PROBLEMS

- Write out the regression equations in score form for the correlation table in problem 2, page 286.
 - Compute $\sigma_{(\text{est } Y)}$ and $\sigma_{(\text{est } X)}$.
 - What is the most probable height of a boy who weighs 30 pounds? 45 pounds? What is the most probable weight of a boy who is 36 in tall? 40 in tall?
- In problem 3, page 286, find the most probable grade made by a child whose score on Army Alpha is 120. What is the $PE_{(\text{est})}$ of this grade?
- What is the most probable algebra grade of a child whose I.Q. is 100 (data from problem 4, p. 287)? What is the $PE_{(\text{est})}$ of this grade?
- Given the following data for two tests:

History(X)	English(Y)
$M_X = 75.00$	$M_Y = 70.00$
$\sigma_x = 6.00$	$\sigma_y = 8.00$
$r_{xy} = .72$	

- Work out the regression equations in score form.
 - Predict the probable grade in English of a student whose history mark is 65. Find the $\sigma_{(\text{est})}$ of this prediction.
 - If r_{xy} had been .84 (σ 's and means remaining the same) how much would $\sigma_{(\text{est } Y)}$ be reduced?
- Plot the regression lines in on the correlation diagram given in problem 4, page 287. Calculate the means of the Y -arrays (successive Y -columns), plot these as points on the diagram, and join these points with straight lines. Plot, also, the means of the X -arrays and join them with straight lines. Compare these two "lines-through-means" with the two fitted regression lines (see Fig. 47, p. 290).
 - In a group of 115 freshmen, the r between reaction time to light and substitution learning is .30. The σ of the reaction times is 20. What would you estimate the correlation between these two tests to be in a group in which the σ of the reaction times is 25?

ANSWERS

1. $\bar{Y} = .40X + 24.12$; $\bar{X} = 1.26Y - 11.52$
 (a) $\sigma_{(\text{est } Y)} = 1.78$; $\sigma_{(\text{est } X)} = 3.16$
 (b) 36.12 inches; 42.12 inches; 33.84 lbs.; 38.88 lbs.
2. 85.2; $PE_{(\text{est } Y)} = 4.7$
3. $\bar{X} = .37Y + 8.16$. When Y (I.Q.) is 100, \bar{X} (algebra) is 45.2.
 $PE_{(\text{est } X)} = 4.6$
4. (a) $\bar{Y} = .96X - 2$; $\bar{X} = .54Y + 37.2$
 (b) 60.4; $\sigma_{(\text{est } Y)} = 5.5$
 (c) 22%
6. $r = .65$

CHAPTER XI

THE RELIABILITY AND VALIDITY OF TEST SCORES

I. RELIABILITY OF TEST SCORES

THE reliability of a test, or of any measuring instrument, depends upon the *consistency* with which it gauges the ability of those to whom it is applied. If the members of a group take a test the second time, and each individual makes a score which differs very little, or not at all, from his first record, the test is said to be reliable. On the other hand, if there are wide discrepancies between the scores made on the first and second administrations of the test, and if such differences occur in large number, scores on the test are inconsistent and unstable, and the test is unreliable.

1. The Determination of Reliability

There are three methods in general use for determining the reliability of a test. These are:

- (1) Repetition of the test
- (2) Use of parallel forms of the test
- (3) Split-half method

Repetition of a test is the simplest method of determining reliability; the test is given, and then repeated, and the correlation calculated between the first and second sets of scores. When parallel forms of the test have been constructed, the correlation between Form A, say, and Form B is taken as a measure of the self-correlation or reliability of the test. This method is employed by the makers of most standard psychological and educational tests, for which parallel forms are usually available. In the split-half method, the test is broken

into two equivalent parts, and the correlation of these half-tests is computed. From the half-test reliability, the self-correlation of the whole test is estimated by the method described on page 318.

These three methods of determining reliability furnish *estimates*, rather than strictly accurate measures, of the consistency of test scores. Each of the methods described is subject to certain qualifications, and each holds good only under certain conditions. If a test is repeated immediately, many subjects will undoubtedly recall their first answers and spend their time on new material, thus increasing their scores. Besides this memory effect, practice and the confidence induced by familiarity with the material will also almost certainly affect scores when one takes a test for the second time. This transfer effect will doubtless be different from person to person. But its net effect is to make for closer agreement between scores obtained on the first and second giving of the test than would otherwise be the case. As a consequence, the reliability coefficient by the repetition method is nearly always too high. When a sufficient time interval has elapsed between the first and second administrations of a test to offset (in large part, at least) memory, practice and other effects, the reliability coefficient will be a close estimate of the actual consistency of the test scores. The reliability of those tests which contain novel features, and which are highly susceptible to practice, will be less accurately determined by the repetition method than will be the reliability of tests involving routine operations little affected by practice. Because of the difficulty in knowing and in controlling the conditions which influence the scores on the second administration of a test, the repetition method of determining test reliability is less generally used than the other two methods.

The use of parallel forms in determining the reliability coefficient is usually satisfactory if sufficient time intervenes between the administration of the two forms to preclude memory and practice effect. When Form B of a test is given to a group shortly after Form A, the scores on the second test will usually

be increased through practice and familiarity. If such increase is approximately constant, however (say 3 to 5 points added to each Form B score), the reliability coefficient of the test will not be affected, since paired A and B scores maintain the same relative positions in the two distributions. When the mean increase due to practice has been determined, a constant amount can be subtracted from Form B scores to make them comparable to Form A scores.* In drawing up alternate forms of a test one should be careful to match test material for content, difficulty and form†; but one must be careful not to make test forms *too* much alike. If parallel forms are practically identical, the reliability coefficient of the test will be too high; on the other hand, if parallel forms are not sufficiently "duplicate," the reliability coefficient will be too low.

The split-half method of measuring reliability is employed when it is not feasible to construct a parallel form of the test nor wise to repeat the test. This situation occurs with many performance tests, as well as with tests and questionnaires dealing with personality traits, attitudes and the like. A performance test (e.g., picture completion, puzzle-solving, form-board) is often a very different test when repeated, as the child is familiar with its procedure and content. Likewise, many personality tests cannot well be given in parallel form, nor repeated, because of radical changes in a subject's attitude and interest when taking such a test the second time.

The chief drawback to the split-half method is that the subject is tested on only one occasion. Relatively temporary influences, therefore (e.g., attitudes, feelings and the like), which would probably differ at another time and thus cancel out, affect the scores on both halves of the test in the same way.

* In the Otis Self-Administering Test of Mental Ability, Higher Examination, for instance, the author suggests that when Form B, which is slightly more difficult than Form A, is given *first*, 4 points be *added* to each person's score. This is to make these scores equivalent to the norms for Form B when this test is given *after* Form A, as it usually is. See *Manual of Directions*, Otis S-A Test, 1922, page 3.

† Ruch, G. M., *The Objective or New-Type Examination*, 1929, Chapter 2, pp. 66-67.

This tends to make the reliability coefficient too high. The longer the test, the less the probability that the effects of temporary and variable disturbances will be cumulative and in one direction, and the more accurate the estimate of reliability.

There are many factors which may affect the reliability of a test besides fluctuations in interest and attention, shifts in emotional attitude, and the differential effects of memory and practice. To these above mentioned "psychological factors" must be added such environmental disturbances as distractions, noises, interruptions, errors in scoring and the like. All of these variable influences (both environmental and psychological) are subsumed under the head of "chance" errors. The cumulative effect of chance errors is assumed to influence a score in such a way as to cause it to vary above — as often as below — its theoretically true value. The reliability coefficient, therefore, is in the nature of a quantitative estimate of the importance of chance or variable errors. Constant errors, in contradistinction to chance errors, work in only *one* direction. Such errors may raise or lower all of the scores on a repeated test or on the parallel form of a test but will not affect the reliability coefficient. To illustrate, if every paper on Form B of a test is scored five points too high, the self-correlation of the test will not be affected (i.e., the correlation between Form A and Form B) but all of the scores on the second form will be in error by five points.

How high should the self-correlation of a test be in order for the reliability of the test to be regarded as satisfactory? This is an important question, and its answer depends upon the nature of the test, the size and variability of the group tested, and the purposes for which the test is intended. To distinguish reliably between the means of two relatively small groups of narrow range of ability (for example, a group of children in the fifth grade, and a group of children in the sixth grade), a reliability coefficient need be no higher than .50 or .60.* If

* Kelley, T. L., *Interpretation of Educational Measurements*, 1927, pp. 210-211.

the test is to be used to differentiate among the scores made by *individuals* in the group, however, its reliability must be .90 or more. Most makers of general intelligence tests, and of educational achievement examinations, report correlations of .90 or more between duplicate forms of their tests in unselected groups of fairly narrow age range. Since the reliability coefficient of a test is directly affected by the range and variability of the test scores (p 322), in reporting a test's reliability the size and the variability of the group upon which the r is based should always be given.

2. The Effect upon Reliability of Lengthening or Repeating a Test

(1) The Reliability Coefficient from Many Applications or Repetitions of a Given Test

We know that the average of five determinations of height, say, is more reliable than a single determination (p. 200), and that, in general, the average of 10 determinations will be more reliable than the average of five. On the same principle, increasing the length of a test, or averaging the results obtained from several applications of a test, or its parallel forms, should tend to increase its reliability. If the self-correlation of a test is not satisfactory what will be the effect of doubling or tripling the test's length? To answer this question experimentally would require considerable time and labor; hence it is fortunate that a good measure of the effect of lengthening or repeating a test may be obtained by use of the Spearman-Brown "prophecy formula" *

$$r_{nn} = \frac{nr_{11}}{1 + (n - 1)r_{11}} \quad (69)$$

(Spearman-Brown formula for estimating the correlation between n forms of a test and n other similar forms)

* Spearman, C., *Correlation Calculated from Faulty Data*, British Journal of Psychology, 1910, 3, p. 281.

Brown, W., *Some Experimental Results in the Correlation of Mental Abilities*, British Journal of Psychology, 1910, 3, p. 299.

in which r_{nn} represents the correlation between n forms of a test and n parallel forms (or the *average* of n forms against the *average* of n other forms) and r_{11} is the reliability coefficient. The subscripts ("11") show that the correlation is between two forms of the *same* test.

To illustrate the use of formula (69), suppose that in a group of 100 adults the self-correlation of a test is .70. What will be the effect upon the test's reliability of tripling the length of the test? Substituting $r = .70$ and $n = 3$ in formula (69) and solving for r_{nn} we have

$$r_{nn} = \frac{3 \times .70}{1 + 2 \times .70} = \frac{2.10}{2.40} = .88$$

Tripling the test's length, therefore, increases its self-correlation from .70 to .88. In lieu of tripling the length of the test, we could give three parallel forms of the test and average the three scores made by each person. The reliability of these mean scores (each based upon three measures) will be the same, as far as purely statistical factors are concerned, as the reliability got by tripling the length of the test.

The probable error* of a reliability coefficient estimated by the Spearman-Brown formula is larger than the *PE* of a computed r (p. 280). The formula is

$$PE_{r_{nn}} = \frac{.6745[n(1 - r_{11}^2)]}{\sqrt{N}[1 + (n - 1)r_{11}]^2} \quad (70)$$

(probable error of the reliability coefficient of a lengthened test,
when the reliability coefficient has been estimated
by the Spearman-Brown formula)

in which r_{11} is the original reliability coefficient, n the number of times the test has been increased, and N is the size of the sample. In the problem above, the *PE* of $r_{nn} = .88$ is .02 for the sample of 100 cases.

* Shen, E., *The Standard Error of Certain Estimated Coefficients of Correlation*, Journal of Educational Psychology, 1924, 15, pp. 462-465.

Shen, E., *A Note on the Standard Error of the Spearman-Brown Formula*, Journal of Educational Psychology, 1926, 17, pp. 93-94.

The prophecy formula may also be used to find how many times a test should be given in order to attain a given standard of reliability. Suppose that the self-correlation of a test is .80. How much will this test have to be lengthened, or how many times repeated, in order to insure a reliability coefficient of .95? Substituting $r_{11} = .80$ and $r_{nn} = .95$ in the formula and solving for n , we have

$$.95 = \frac{80n}{1 + .80n - 80} = \frac{.80n}{.20 + .80n}$$

and $n = 4.75$ or 5 in whole numbers.

The test must be five times its present length, therefore, or five parallel forms must be given and averaged, before the self-correlation of the test will reach .95.

Predictions of test reliability by the prophecy formula are valid only when the items or questions added to the test cover the same ground, are of equal range of difficulty, and are as reliable as the items of the original test. When these conditions are satisfied there is no reason, as far as the mathematics of the process is concerned, why we could not raise the self-correlation of a test to any desired figure, simply by continuing to increase its length, or by continuing to repeat it. It is highly improbable, however, that the reliability coefficient of a test could be increased indefinitely. In the first place, it is impracticable, if not impossible, to increase a test's length, say, 10 to 15 times. Furthermore, beyond a certain point, boredom, fatigue, loss of incentive and the like will inevitably affect our results, and lead to "diminishing returns." When the material added to the test is carefully selected and is strictly comparable to the original test items and when the motivation remains substantially constant, the experimental evidence* indicates that

* Holzinger, K. J., and Clayton, B., *Further Experiments in the Application of Spearman's Prophecy Formula*, Journal of Educational Psychology, 1925, 16, pp. 289-299

Ruch, G. M., Ackerson, Luton, and Jackson, J. D., *An Empirical Study of the Spearman-Brown Formula as Applied to Educational Test Material*, Journal of Educational Psychology, 1926, 17, pp. 309-313

a test may be increased to six or seven times its original length, and the prophecy formula will still give a close estimate of empirically determined results. After the first four or five lengthenings the prophecy formula may "over-predict," give higher estimated reliabilities than those obtained by actual calculation. But this is not a particularly serious drawback to the use of the formula, as a test which needs so much lengthening in order to be reliable should either be radically changed in form or content, or, better still perhaps, discarded in favor of a better test.

The prophecy formula may be applied to ratings, estimates, and other judgments, as well as to test items. When one measures the reliability of a personality rating scale, for instance, by correlating the ratings made by two equally competent judges, he may employ the prophecy formula to estimate the increased reliability in the ratings which might be expected if there were four, six, or more judges.*

(2) The Reliability Coefficient from One Application of a Test

When a test has no parallel form and cannot well be repeated, we may calculate the reliability of *half* of the test and then proceed to estimate the reliability of the *whole* test by means of the Spearman-Brown formula. This method is sometimes called the "split-half technique." The procedure is as follows: First, we make up two sets of scores by combining, say, alternate exercises or items in the test. The first set of scores represents performance on the odd-numbered items, 1, 3, 5, 7, etc.; and the second set performance on the even-numbered items, 2, 4, 6, 8, etc. Other ways of making the two halves of the test as comparable as possible in content, difficulty and susceptibility to practice, may be employed, but the method described is the one most commonly used. From the self-

* Clark, E. L., *Spearman-Brown Formula Applied to Ratings of Personality Traits*, Journal of Educational Psychology, 1935, 26, pp. 552-555.

Remmers, H. H., Shock, N. W., and Kelly, E. L., *An Empirical Study of the Validity of the Spearman-Brown Formula as Applied to the Purdue Rating Scale*, Journal of Educational Psychology, 1927, 18, pp. 187-195.

correlation of the half-test, the reliability coefficient of the whole test may be estimated by the formula

$$r_{11} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1 + r_{\frac{1}{2}\frac{1}{2}}} \quad (71)$$

(Spearman-Brown formula for estimating reliability from two comparable halves of a test)

in which r_{11} is the reliability coefficient of the whole test, and $r_{\frac{1}{2}\frac{1}{2}}$, found experimentally, is the reliability coefficient of one-half of the test. When the reliability coefficient of one-half of a test ($r_{\frac{1}{2}\frac{1}{2}}$) is .60, it follows from formula (71) that the reliability of the whole test (r_{11}) will be .75.

3. The Index of Reliability

By an individual's "true score" on a test (p. 198) is meant the mean of a very large number of measures made of the given individual on the same test or parallel forms of the test administered under approximately identical conditions. The correlation between a series of obtained scores and their corresponding "theoretically true" scores may be found by the formula

$$r_{1\infty} = \sqrt{r_{11}} \quad (72)$$

(correlation between obtained scores on a given test and true scores in the function measured by the test)

in which r_{11} is the reliability coefficient of the given test, and $r_{1\infty}$ represents the correlation between obtained and true scores. The symbol " ∞ " (infinity) will be used hereafter to designate true scores, that is, those scores gotten from an "infinite" number of administrations of the test to the same group.

The coefficient $r_{1\infty}$ is often called the *index of reliability* — it measures the reliability or consistency of a test by showing how well the obtained test scores agree with their theoretically true counterparts. The index of reliability also gives the maximum correlation which the given test is capable of yielding.

This follows from the fact that "the highest possible correlation which can be obtained (except as chance might occasionally lead to higher spurious correlation) between a test and a second measure is with that which truly represents what the test actually measures, that is, the correlation between the test and the true scores of individuals in just such tests." *

To illustrate the application of the index of reliability, suppose that for a given test the self-correlation is .64. Then $r_{1\infty} = \sqrt{.64}$ or .80; and .80 is the highest correlation of which this test is capable, since it represents the relationship between obtained test scores and true test scores in the same function. If the self-correlation of a test is only .25, so that $r_{1\infty} = \sqrt{.25}$ or .50, it is obviously a waste of time to continue using this test without lengthening or otherwise improving it. A test whose index of reliability is only .50 is an extremely poor estimate of the function which it is trying to measure.

4. The Standard and Probable Errors of an Obtained Score

The effect of variable or chance errors in producing divergences of an obtained score from its theoretically true value may be measured by the formula

$$\sigma_{1\infty} = \sigma_1 \sqrt{1 - r_{11}} \quad (73)$$

(standard error of an obtained score (standard error of measurement) when the σ 's of the test and its alternate form are equal)

in which $\sigma_{1\infty}$ is the standard error of an obtained score (sometimes called the "standard error of measurement"), σ_1 is the standard deviation of the test scores and r_{11} is the reliability coefficient of the test. The subscript " $_{1\infty}$ " indicates this standard deviation to be a measure of the error made in taking an obtained score (i.e., 1) as an estimate of its true value (i.e., ∞).

Instead of $\sigma_{1\infty}$, the probable error of an obtained score,

* Kelley, T. L., *The Reliability of Test Scores*, Journal of Educational Research, 1921, 3, p. 327.

$PE_{1\infty}$, which is perhaps more often used, may be calculated from the formula

$$PE_{1\infty} = .6745\sigma_1\sqrt{1 - r_{11}} \quad (74)$$

(*PE of an obtained score (PE of measurement) when the σ 's of the test and its alternate form are equal*)

To illustrate $PE_{1\infty}$ suppose that in a group of 300 college freshmen, the reliability coefficient of a general intelligence test is .92, and the σ of the distribution is 15. From formula (74) we have

$$PE_{1\infty} = .6745 \times 15\sqrt{1 - .92} = 2.86 \text{ or } 3 \text{ in whole numbers.}$$

This result may be interpreted to mean that the chances are *even* (50 in 100) that the obtained score of *any* individual in the group of 300 does not differ from his true score by more than ± 3 points. If a subject has a score of 85, therefore, the chances are even that his true score falls within the limits 85 ± 3 or between 82 and 88. We may be sure that his true score falls between 73 and 97 ($\pm 4PE$). Generalizing for the whole group, we should expect about half of the 300 scores to be in error (as compared with their true scores) by less than 3 points, and the other half to be in error by 3 or more than 3 points.

In formulas (73) and (74) the σ 's of the test and its parallel form are assumed to be equal. When this is not true nor approximately true, we must use the following formulas:

$$\sigma_{1\infty} = \frac{(\sigma_1 + \sigma_2)}{2} \sqrt{1 - r_{11}} \quad (75)$$

and

$$PE_{1\infty} = .6745 \left[\frac{(\sigma_1 + \sigma_2)}{2} \sqrt{1 - r_{11}} \right] \quad (76)$$

(*standard error and probable error of an obtained score, the σ 's of the test and its alternate form not assumed to be equal*)

In our illustration above, if the σ 's of the two forms of our test had been 15 and 10, respectively, $PE_{1\infty}$ would have been

$$PE_{1\infty} = .6745 \times 12.5 \sqrt{1 - .92} = 2.39 \text{ or } 2 \text{ in whole numbers.}$$

The student must be careful not to confuse the formulas for $\sigma_{(\text{est})}$ and $PE_{(\text{est})}$ (see p. 300) with those for $\sigma_{1\infty}$ and $PE_{1\infty}$. The "estimate" formulas enable us to say with what degree of assurance we can predict an individual's score on one test, when we know his score on a second (and usually a different) test. The actual prediction of the most probable score is made, of course, by means of the regression equation connecting the two variables (p. 297). The $\sigma_{1\infty}$ and $PE_{1\infty}$ formulas, on the other hand, deal with the relation between actual scores and the theoretically true scores which they represent more or less well. Hence they are essentially measures of reliability — of how well obtained measures represent true measures of the same function.

When tests are scored in different units, the $\sigma_{1\infty}$ of one test cannot be compared directly with the $\sigma_{1\infty}$ of another test. The reliability of the score made by a child on a reading test, for example, cannot be compared directly with the reliability of his score on a general intelligence test. One method of making a comparison in such cases is to employ a ratio similar to the coefficient of variation, V (p. 51). If the $\frac{\sigma_{A\infty}}{M_A}$ of Test A is smaller than the $\frac{\sigma_{B\infty}}{M_B}$ of Test B, the scores on the first test may be regarded as the more reliable. To illustrate, suppose that in a given group the $\sigma_{1\infty}$ of a digit-span test is .35, and the test mean is 8; while in the same group the $\sigma_{2\infty}$ of a general intelligence test is 4 and the test mean is 150. The $\frac{\sigma_{1\infty}}{M_1}$ of the first test is .044, and the $\frac{\sigma_{2\infty}}{M_2}$ of the second test is .027; hence the second test is the more reliable in the given range.

5. The Dependence of the Reliability Coefficient upon the Size and Variability of the Group

The reliability coefficient of a test administered to a small group, a single grade, say, cannot be directly compared with

the reliability coefficient of the same test when administered to a larger group, e.g., to the children in several grades. This is true because the reliability coefficient (like any correlation coefficient) is affected by the variability of the group; and the larger and more heterogeneous the group, the greater the variability tends to be. If we know the reliability of a test in a narrow range (ordinarily a small group) we can estimate the reliability of the test in an increased range (ordinarily a larger group) by the following formula

$$\frac{\sigma_s}{\sigma_l} = \frac{\sqrt{1 - r_{ss}}}{\sqrt{1 - r_{ll}}} \quad (77)$$

(relation between σ 's and reliability coefficients obtained in different ranges when the test is equally effective throughout both ranges)

in which σ_s and σ_l are the σ 's of the test scores in the small and large group, respectively, and r_{ss} and r_{ll} are the reliability coefficients in the small and large groups.

To illustrate formula (77), suppose that in a single fifth grade $r_{ss} = .50$, and $\sigma_s = 5.00$; and that in a larger group made up of children from grades three to seven, $\sigma_l = 15.00$. Assuming the test to be as effective in the large group as in the small, what is the reliability coefficient of the test in the large group? If we substitute for σ_s , σ_l , and r_{ss} , in formula (77), $r_{ll} = .94$. This means that a reliability coefficient of .50 in the small group indicates as great test consistency as a reliability coefficient of .94 in a group in which the range of talent is three times as great. Formula (77) may be employed to determine whether a test is equally effective in different parts of the range (r_{ss}) as in the entire range (r_{ll}); or in one range as in another.

Even when one test is "intrinsically" as good as another, as far as range of difficulty and content are concerned, differences in reliability, as between the two tests, will still arise from differences in the lengths of the tests, and from differences in the heterogeneity of the groups to whom the tests have been administered. We cannot, therefore, compare the effectiveness

of the same test in different groups or of different tests in the same group unless we take account of these factors. Suppose, for example, that the reliability of Test *A* is .85 in a certain group, and that the reliability of Test *B* is .70 in the same group. If the time limit of Test *A* is twice the time limit of Test *B* (i.e., Test *A* is about twice as long as Test *B*), we may predict from formula (69) that the reliability of Test *B* would be .82 if it were sufficiently lengthened to require the same time as Test *A*. Again, in a group the σ of which is twice as great as the σ of the given group, the reliability of Test *B* would be $\frac{1}{2} = \frac{\sqrt{1 - r_{11}}}{\sqrt{1 - .70}}$ or .93. These examples make clear the need

for stating (1) the length of the test, and the time it takes, as well as (2) the size and spread (σ) of the group, in comparing and interpreting reliability coefficients.

II. VALIDITY OF TEST SCORES

The validity of a test, or other measuring instrument, depends upon the *fidelity* with which it measures whatever it purports to measure. A yard-stick is valid when measurements made by it can be checked by other measuring rods. And in the same way, a test is valid when the capacity which it gauges corresponds to the same capacity as otherwise objectively measured and defined. The difference between validity and reliability can be made clear, perhaps, by an illustration. Suppose a clock is set forward 20 minutes. If the clock is a good timepiece, the "time it tells" will be reliable (i.e., consistent), but it will not be *valid* as judged by "standard time."

1. The Determination of Validity through Correlation with a Criterion

The validity of a test is determined directly, whenever possible, by finding the correlation between the test and some independent criterion. A criterion is an objective measure in terms of which the value of the test is estimated or judged. The

criteria of a general intelligence examination, for example, may be school marks, ratings for intelligence, or some other test believed to be valid, such as Stanford-Binet.* The criterion of a trade test is demonstrated ability in the trade as shown by actual performance † A high correlation between a test and its criterion may be taken as evidence of validity, provided both the test and its criterion are reliable. But before accepting criterion-correlations as final, we should know the reliability of the test, and if possible the reliability of the criterion.‡

2. The Effect upon Validity of Lengthening a Test

Just as lengthening a test increases its reliability, so also lengthening a test increases its validity by making the test a better measure of some criterion. The effect upon validity of increasing the length of a test may be measured by the following formula

$$r_{c(n x_1)} = \frac{n r_{c x_1}}{\sqrt{n + n(n-1) r_{x_1 x_1}}} \quad (78)$$

(correlation between a criterion and (1) a test lengthened n times or (2) the average of n parallel forms of a test) §

in which $r_{c(n x_1)}$ = the correlation between the criterion (c) and n forms of test X_1 , or test X_1 lengthened n times.

$r_{c x_1}$ = the correlation between the criterion (c) and the given test X_1 .

$r_{x_1 x_1}$ = the reliability coefficient of the test X_1 .

n = number of parallel forms of test X_1 , or the number of times it is lengthened.

To illustrate formula (78), suppose Test A has a reliability coefficient (r_{11}) of .70 and a correlation of .40 (r_{ca}) with a criterion (c). What would be the correlation of Test A with the same criterion (i.e., its validity coefficient) if the test were

* Yoakum, C S, Yerkes, R. M, *Army Mental Tests*, 1920, Chapter 1

† Bingham, W V., *Aptitudes and Aptitude Testing*, 1937, Chapter XVI.

‡ Paterson, D G. et al., *Minnesota Mechanical Ability Tests*, 1930, Chapters 3, 6, 7, 9.

§ Kelley, T. L., *Statistical Method*, 1923, p. 199.

tripled in length? Substituting $r_{x_1x_1} = .70$, $r_{cx_1} = .40$, and $n = 3$, in formula (78)

$$r_{c(3A)} = \frac{3 \times .40}{\sqrt{3 + 3 \times 2 \times .70}} = .45$$

Hence, (1) tripling the length of Test *A*, or (2) giving three forms of the test and averaging the three scores for each subject, increases from .40 to .45 the test's correlation with its criterion. It should be noted that tripling the length of the test will also increase its reliability from .70 to .88 [Spearman-Brown formula (69), p. 315] This increase in reliability with increase in validity emphasizes the close relationship between these two measures of a test's efficiency.

By rearranging formula (78) so that n is on the left hand side of the equation, we can find how many forms of a test would have to be administered, or how many times the test would have to be lengthened, in order for a given validity coefficient to be attained. The formula becomes

$$n = \frac{r_{c(nx_1)}^2(1 - r_{x_1x_1})}{r_{cx_1}^2 - r_{c(nx_1)}^2 \cdot r_{x_1x_1}} \quad (79)$$

(amount by which a test must be lengthened (n), or number of parallel forms which must be averaged, in order to give a specified validity coefficient)

Suppose that a trade Test *T* in a small group has a reliability coefficient (r_{11}) of .50, and a correlation with its criterion (r_{cx_1}) of .30. How many times would the test have to be lengthened in order to yield a validity coefficient ($r_{c(nx_1)}$) of .40? Substituting $r_{c(nx_1)} = .40$, $r_{x_1x_1} = .50$, $r_{cx_1} = .30$ in formula (79) and solving for n , we have

$$n = \frac{.16 \times .50}{.09 - .16 \times .50} = 8$$

Test *T* must be 8 times its present length, therefore, before its validity coefficient will increase from .30 to .40. Note, also, that increasing Test *T* to 8 times its present length raises its reliability coefficient from .50 to .89 (p. 315).

A little experimentation with formula (79) will show that, if one sets his standard of validity for a given test too high, the formula will give negative (i.e., impossible) values. This follows from the fact that the highest *possible* correlation which a test can have with a given criterion is limited by the reliability coefficient of the test. The upper limit of a test's validity is given by the formula

$$r_{c(\infty)} = \frac{r_{cx_1}}{\sqrt{r_{x_1x_1}}} \quad (80)$$

(correlation between a criterion and "true" scores in a given test)

in which $r_{c(\infty)}$ is the correlation between the criterion (c) and true scores X_∞ in the test X_1 ; r_{cx_1} is the correlation of the test and the criterion (c), and $r_{x_1x_1}$ is the reliability coefficient of the test X_1 .

In the example above,

$$r_{c(\infty)} = \frac{.30}{\sqrt{.50}} = .42$$

and the trade Test T cannot be expected to return a validity coefficient higher than .42 no matter how much it is lengthened.

3. "Indirect" Measures of Validity

When a reliable criterion is not available, there are several indirect methods which may be employed in determining the validity of a test. (1) One method is to compute the average correlation which each test in a battery shows with all of the other tests, and to estimate the validity (i.e., the representativeness) of each test by the size of its average correlation. Again, following essentially the same method, one may combine the scores on a number of tests designed to measure the same function (memory, say) and consider as most valid that test which correlates highest with the average of them all. Whitley,* for example, working with three "discrimination" tests — naming colors, naming forms, and naming objects —

* Whitley, Mary T., *An Empirical Study of Certain Tests for Individual Differences*, Archives of Psychology, 1911, 19, p. 78.

obtained the following correlations between each test and the average of all three:

$$\text{Average of all three tests and } \left\{ \begin{array}{l} \text{Naming colors } r = .67 \\ \text{Naming forms } r = .99 \\ \text{Naming objects } r = .96 \end{array} \right.$$

She concludes that "Naming forms seems more a typical test in so far as it measures an ability common to all these three tests." Anastasi* found that, of eight tests of immediate memory, the paired associates test (geometrical forms paired against numbers) had the largest average correlation (i.e., .49) with the other tests of the battery. In a sense, then, this test is the most valid measure of the function tapped in common by all of the tests.

(2) A second indirect method of validating a test is to rely upon competent judgments as to the suitability (validity) of the material included. This method is generally employed in personality questionnaires and in inventories dealing with neurotic symptoms, introversion-extroversion, attitudes, interests, and the like.† An objective criterion of validity called the "criterion of internal consistency" has been proposed for questionnaire material by the Thurstones‡ This criterion admits in the final questionnaire only those items or questions which in preliminary experiments have been found to distinguish between the high-scoring and the low-scoring members of the group. The items in the final test, therefore, "hang together" in the sense that they all work in the same direction and presumably measure the same common trait.

The validity of most standard tests of educational achievement usually depends upon the consensus of opinion of teachers and other competent judges as to the adequacy of the material included. Courses of study, requirements for the different

* Anastasi, A., *A Group Factor in Immediate Memory*, Archives of Psychology, 1930, 120, p. 41.

† Garrett, H. E., and Schneek, M. R., *Psychological Tests, Methods, and Results*, 1933, Chapter 3, Part II, pp 144-150.

‡ Thurstone, L. L., and Thurstone, T. G., *A Neurotic Inventory*, Journal of Social Psychology, 1930, 1, pp. 3-30.

grades, curricula from different sections of the country, are all carefully culled over by the test-makers, to determine what material in history, English, arithmetic, etc., should be included in an educational test battery designed, say, for the seventh and eighth grades.* In its final form, therefore, the educational achievement test represents carefully selected items drawn from all available sources of information.

(3) Still other methods of determining validity may be employed in lieu of objective criteria. The *index of reliability*, it will be recalled (p. 319), gives the correlation between obtained scores and theoretically true scores in the same function. The *index*, therefore, is in reality a validity coefficient, since it tells us how well our test is measuring true ability in the function which it purports to measure. Hartshorne and May † have used the index of reliability in determining the theoretical validity of certain of their deception tests.

4. The Relation between Validity and Reliability ‡

From the preceding sections, the reader has doubtless become aware of the close connection between reliability and validity. The real distinction between the two concepts is one of emphasis. Both stress test efficiency. Reliability, however, does not go beyond the test measures themselves; while validity implies evaluation in terms of *outside* — and independent — criteria. It is not an especially formidable task to measure the reliability of a test by the methods described in this chapter. But difficulty is encountered when we try to obtain really authentic criteria in terms of which to validate our test. Criteria

* Ruch, G. M., *The Objective or New-Type Examination*, 1929, Chapter 2.

† Hartshorne, H., and May, M. A., *Studies in Deceit*, 1928, Part II, p. 126.

See also: — Kelley, T. L., and Shen, E., *The Statistical Treatment of Certain Typical Problems*, Chapter 23, in *Foundations of Experimental Psychology*, 1929.

‡ For further discussion of reliability and validity, see: —

Willoughby, R. R., *The Concept of Reliability*, *Psychological Review*, 1935, 42, pp. 153-165.

Dunlap, J. W., *Comparable Tests and Reliability*, *Journal of Educational Psychology*, 1933, 24, pp. 442-453.

scores for mental tests can constitute, at best, only approximate measures; for when accessible and highly reliable criteria exist, we take these criterion scores and do not bother with the test. Perhaps it will be helpful to consider here the distinction between physical and mental measurements with respect to reliability and validity. The reliability of the measurements made by scales, thermometers, yard-sticks, chronoscopes, clocks, etc., is determined by making repeated measurements of the same facts; and validity is determined by comparing the measures returned by the given instrument with highly precise (if arbitrary) "standard" measures. The reliability of mental measures is found in the same way. But since precise and independent "standards" (criteria) do not exist in mental measurement, the validity of a test can never be estimated as precisely as can the validity of a thermometer or a rheostat.

In order to be valid a test must be reliable, and the more reliable the test the higher its validity coefficient as formula (78) indicates. A highly reliable test is *always* valid theoretically. Thus, if a test has a reliability coefficient of .90, its index of reliability is $\sqrt{.90}$ or .95, and the obtained test scores correlate .95 with theoretically true scores in the function. A test may be theoretically valid, however, and at the same time exhibit so little correlation with independent criteria as to have little *practical* validity. Moreover, if the test has no correlation with other psychological tests, and hence measures nothing that we can identify, it has little value of any sort. The simple tapping test is one example among many others. This test can be made highly reliable by increasing its length, so that its index of reliability (theoretical validity) will be high. But this test rarely correlates highly enough with other tests to make it a useful measure of any function.

We may summarize this discussion of reliability and validity as follows:

- (1) The two concepts, reliability and validity, refer to different aspects of what is essentially the same thing, namely, test efficiency.

- (2) A reliable test *cannot* be invalid, *theoretically*, but it may be *practically* invalid as judged by its correlations with such independent criteria as we possess.
- (3) Strictly speaking, a valid test *cannot* be unreliable, since the correlation of a test with a criterion is limited by its own reliability coefficient.

III. THE ESTIMATION OF TRUE MEASURES

1. The Estimation of the True σ of a Test

Chance or variable errors have a marked effect upon the standard deviation of a test, and it is often important to estimate just how extensive this effect is. The relation of the σ , calculated from obtained scores on a test, to the σ of true scores on the same test is given by the formula

$$\sigma_{\infty} = \sigma_1 \sqrt{r_{11}} \quad (81)$$

(relation between true and obtained σ 's of a set of test scores)

in which σ_{∞} is the σ of the true test scores, σ_1 is the σ of the obtained test scores, and r_{11} is the reliability coefficient of the test.

Suppose an educational achievement test has been administered to 50 children. The obtained standard deviation, σ_1 , is 10, and the reliability coefficient of the test (r_{11}) is .50. What is σ_{∞} , the σ of the true scores from which variable or accidental errors have been eliminated? Substituting $\sigma_1 = 10$, and $r_{11} = .50$ in formula (81)

$$\begin{aligned} \sigma_{\infty} &= 10\sqrt{.50} \\ &= 7.07 \end{aligned}$$

and the "true σ " of the test is about 7 points.

It is clear from (81) that σ_{∞} will *always* be smaller than σ_1 , except in the improbable event in which $r_{11} = 1.00$. The effect of chance errors of measurement, then, is *always* to increase the spread (σ_1) of obtained test scores.

At several places in Chapter VIII (pp. 202 and 244), the point was stressed that the standard error of the mean

$\left[\frac{\sigma}{\sqrt{N}} \right]$ measures chance or accidental errors of measurement as well as "errors" or fluctuations arising from sampling. It may now be shown how we can distinguish between these two sources of error, and measure their effects separately. Let M_1 and σ_1 represent the mean and the σ of Test I. Then replacing σ_1 by σ_∞ , we have

$$\sigma_{M_1} (\text{sampling}) = \frac{\sigma_\infty}{\sqrt{N}} = \frac{\sigma_1 \sqrt{r_{11}}}{\sqrt{N}} \quad (82)$$

as the standard error of the mean due to fluctuations ("errors") of sampling *only*. Furthermore,

$$\sigma_{M_1} (\text{chance}) = \frac{\sigma_1 \sqrt{1 - r_{11}}}{\sqrt{N}} \quad (83)$$

is the standard error of the mean due to chance errors of measurement *only*.

In the example, given above, of the educational achievement test administered to 50 children, the obtained $\sigma(\sigma_1)$ was 10 and the reliability coefficient (r_{11}) was .50. Hence the "total" standard error of the mean of this test is

$$\sigma_{M_1} = \frac{10}{\sqrt{50}} = 1.41$$

The standard error of the mean attributable to *sampling errors only* is

$$\sigma_{M_1} (\text{sampling}) = \frac{10\sqrt{.50}}{\sqrt{50}} = 1.00$$

and the standard error of the mean due to *chance errors only* is

$$\sigma_{M_1} (\text{chance}) = \frac{10\sqrt{1 - .50}}{\sqrt{50}} = 1.00$$

Ordinarily, the "total" standard error of the mean is the value wanted, since it takes account of *all* of the variable factors which lead to unreliability in the mean.*

* Note that $\sigma_{M_1}^2 = \sigma_{M_1}^2 (\text{sampling}) + \sigma_{M_1}^2 (\text{chance})$.

2. The Prediction of True Scores and their Standard Error

In many studies in which mental and educational tests are employed, it is useful to be able to estimate a subject's *true score* in a given test, and to know the standard error of such an estimated true value. A true score in a test may be estimated from the following formula

$$\bar{X}_{\infty} = r_{11}(X_1 - M_{x_1}) + M_{x_1} \quad (84)$$

(estimate of a true measure from an obtained measure of the same function)

in which

\bar{X}_{∞} = estimated true score

X_1 = obtained score in the test

M_{x_1} = mean of the obtained test scores

r_{11} = reliability coefficient of the test.

The standard error of such an estimated true score is

$$\sigma_{\infty,1} = \sigma_1 \sqrt{r_{11} - r_{11}^2} \quad (85)$$

(standard error made in estimating a true score from an obtained score in the same function)

in which $\sigma_{\infty,1}$ is the standard error of a true score (i.e., \bar{X}_{∞}) estimated from an obtained score (i.e., X_1); σ_1 is the obtained σ of the given test; and r_{11} is the reliability coefficient of the test.

In terms of standard or z -scores, formula (84) becomes

$$\bar{z}_{\infty} = r_{11}z_1 \quad (86)$$

(estimate of a true measure from an obtained measure in the same function — standard scores)

in which \bar{z}_{∞} is the estimated true standard score, z_1 is the obtained standard score, and r_{11} is the reliability coefficient of the test. M_{x_1} , of course, is zero. Formula (85), in terms of z -scores, becomes

$$\sigma_{\infty,1} = \sqrt{r_{11} - r_{11}^2} \quad (87)$$

(standard error made in estimating a true score from an obtained score in the same function — standard scores)

since $\sigma_1 = 1.00$.

To illustrate formulas (84) and (85), suppose that the Thorndike Intelligence Examination for High School Graduates has been administered to 400 candidates for college entrance. The mean score is 80, $\sigma_1 = 10$, and $r_{11} = .85$. What is the estimated true score of a candidate who makes a score of 70, and what is the standard error of this estimated true score? Substituting $M = 80$, $\sigma_1 = 10$, $r_{11} = .85$, and $X_1 = 70$ in formulas (84) and (85), we have

$$\bar{X}_\infty = .85(70 - 80) + 80 = 71.5$$

and

$$\sigma_{\infty,1} = 10\sqrt{.85 - .85^2} = 3.6$$

The *estimated* true score of the given candidate is 71.5 with a standard error of 3.6. There are 68 chances in 100, therefore, that this candidate's *actual* true score does not diverge from its *estimated* (or predicted) value (namely, 71.5) by more than ± 3.6 . Or the chances are 68 in 100 that the actual true score lies between 67.9 and 75.1 (between 68 and 75, approximately).

In the above problem, the candidate's standard score is -1.00 [i.e., $\frac{(70 - 80)}{10}$]. Hence his estimated true standard score is

$$\bar{z}_\infty = .85 \times -1 = -.85$$

and its standard error is

$$\sigma_{\infty,1} = \sqrt{.85 - .72} = .36$$

There are 68 chances in 100, therefore, that our candidate's actual true score lies between -1.21 and $-.49$ ($\pm 1\sigma$), when measures are expressed in standard scores.

3. The Correction of a Coefficient of Correlation for Attenuation

We have found (p. 327) that the correlation between a test and its criterion is lowered by the unreliability of the test, and can be raised by increasing the test's reliability. In the same way, the correlation between a test and its criterion is lowered

by the unreliability of the *criterion* and can be raised by increasing the criterion's reliability. More generally, the correlation between *any* two tests is affected directly by the lack of reliability in the *two* tests concerned. In order to estimate the correlation between *true* scores in two tests, therefore, we need a correction which will take account of the unreliability in both series of measurements. Such a correction is given by the formula

$$r_{\infty\infty} = \frac{r_{12}}{\sqrt{r_{11} \cdot r_{22}}} \quad (88)$$

(correlation between true measures in Tests 1 and 2)

in which

$r_{\infty\infty}$ = correlation between true scores in Tests 1 and 2

r_{12} = correlation between obtained scores in Tests 1 and 2

r_{11} = reliability coefficient of Test 1

r_{22} = reliability coefficient of Test 2

Formula (88) is Spearman's well-known "correction for attenuation" formula.* It provides a correction for the effects of those chance or accidental errors in the two tests which lower the reliability coefficients of both tests and thus affect the correlation between them. To illustrate the application of formula (88), let the obtained correlation between two tests *A* and *B* be .60, the reliability coefficient of Test *A* be .80 (r_{11}) and the reliability coefficient of Test *B* be .90 (r_{22}). What is the correlation between Tests *A* and *B*, freed of chance or variable errors? Substituting the given values in formula (88)

$$r_{\infty\infty} = \frac{.60}{\sqrt{.80 \times .90}} = .71$$

which is the estimated correlation between true scores in *A* and *B*. This coefficient represents the "corrected" correlation to be

* Spearman, C., *The Proof and Measurement of Association between Two Things*, American Journal of Psychology, 1904, 15, pp. 72-101.

Spearman, C., *Demonstration of Formulae for True Measurement of Correlation*, American Journal of Psychology, 1907, 18, pp. 161-169.

expected between the two tests when their reliability coefficients each equals 1.00.

It is clear from formula (88) that correcting for chance or variable errors will *always* raise the correlation between two tests unless the reliability coefficients are both 1.00. Chance or variable errors, therefore, always serve to lower or "attenuate" an obtained correlation coefficient. The expression $\sqrt{r_{11}r_{22}}$ sets an upper limit to the correlation which we can obtain between two tests as they stand. In the example above $\sqrt{.80 \times .90} = .85$; hence, Tests *A* and *B* cannot correlate higher than .85, since otherwise their corrected r ($r_{\infty\infty}$) would be greater than 1.00.

Referring again to page 325, the student will note that formula (78) gives the increase to be expected in the validity coefficient of a test when the reliability coefficient of the test is improved. In this case, the criterion is either assumed to be completely reliable, or its reliability is ignored, because unknown. If the reliability of the criterion is known, however, it becomes possible by formula (88) to estimate the increase in the validity of a test when *both* the unreliability of the test and the unreliability of the criterion are considered. Hence, by means of formula (88) we are able to estimate the correlation to be expected between *true* test scores and *true* criterion scores.

4. Assumptions Underlying the Correction for Attenuation

Certain precautions must be observed when applying formula (88), and the student should have clearly in mind what the formula does and upon what assumptions it rests. (1) In the first place, the correction for attenuation assumes zero correlation between chance errors in each test and true scores in the test itself; and it also assumes zero correlation between chance errors in the two tests. These assumptions are probably true (or approximately true) in those cases wherein standard psychological and educational tests are applied to large groups; but there are laboratory situations in which they are exceedingly

doubtful. In judging comparative brightnesses, the lengths of lines, and in reaction time to light and sound, for example, in successive experiments not greatly separated in time, it is probable that many presumably "accidental variations" are carried over (in the nature of a "mental set") from one experimental situation to another.* In such cases formula (88) can be expected to give only approximate corrections. Correction for attenuation is most valid when the interval between the original test and the parallel form of the test is long enough for variations to be truly accidental (p. 312).

(2) The investigator should be careful how he applies formula (88) to correlations which have been averaged, or to tests which have been combined in various ways. In such cases the reliability coefficients may be less than the correlation between the two tests, and r_{∞} greater than 1.00. Such a result is both logically and psychologically meaningless, since it implies that one test is a better measure of another (and presumably *different* trait) than it is of the function which it measures itself. However, if the corrected r is not more than $\pm 4PE$ removed from 1.00,† corrected r 's which are slightly greater than 1.00 may, perhaps, be safely taken to mean that the true r is 1.00.

(3) The correction for attenuation is extremely valuable in giving the "theoretically maximum" correlation which one can expect to obtain between two tests when each is perfectly reliable. Suppose, for example, that the correlation between first year college grades and general intelligence is .46; that the reliability of the general intelligence test is .82, and the reliability of the grades .70. Then the maximum correlation

* Brown, W., and Thomson, G., *The Essentials of Mental Measurement*, 1925, pp. 156-160.

† Thorndike, E. L., et al., *The Measurement of Intelligence*, 1927, pp. 556-564.

For a discussion of the " PE of a corrected r " see Shen, E., *The Standard Error of Certain Estimated Coefficients of Correlation*, Journal of Educational Psychology, 1924, 15, pp. 462-465.

Cureton, E., and Dunlap, J. W., *Spearman's Correction for Attenuation and Its Probable Error*, American Journal of Psychology, 1930, 42, pp. 235-245.

which we could hope to obtain if both measures were perfectly reliable is $\frac{.46}{\sqrt{.70 \times .82}}$ or .60. Knowing that the correlation

between grades and general intelligence, corrected for errors of measurement, has a probable maximum value of .60 gives us a better notion of the "intrinsic" relationship between the two variables. At the same time, the investigator should remember that an r_{∞} of .60 is a theoretical, not an obtained, value; and that it gives an estimate of the relation to be expected when the tests are more effective than they actually were in the present instance. An obtained correlation, after all, describes the situation as it *actually* existed when the two tests were administered. If many sources of error are present so that considerable correction is necessary, it would be better experimental technique to improve the tests and the experimental conditions, than to correct the obtained r in order to approximate an ideal situation which did not exist in the given experimental set-up.

PROBLEMS

1. The reliability coefficient of a test is .60.
 - (a) How much must this test be lengthened in order to raise the self-correlation to .90?
 - (b) What effect will doubling the test's length have upon its reliability coefficient? tripling the test's length?
 - (c) Compute the PE 's of the estimated reliability coefficients if N is 100.
2. A test of 50 items has a reliability coefficient of .78. What is the reliability coefficient
 - (a) of a test having 100 items comparable to the items in the given test?
 - (b) of a test having 125 comparable items?
 - (c) If N is 100 what are the PE 's of the r 's estimated in (a) and (b) above?

3. A given test has a reliability coefficient of .80 and a σ of 20.
 - (a) What is the maximum correlation which this test is capable of yielding as it stands (see p. 319)?
 - (b) What is the standard error of a score obtained on this test?
 - (c) What is the estimated reliability coefficient of this test in a group in which the σ is 15?
4. A test is given to a group of 225 subjects with the following results: $M = 62.50$, $\sigma = 9.62$; $r_{11} = .91$.
 - (a) What is the estimated "true σ " of this test?
 - (b) Compute the standard error of the mean due to sampling fluctuations; and the standard error of the mean due to chance errors.
 - (c) If subject A makes a score of 55 on this test, what is his estimated true score and its standard error ($\sigma_{\infty.1}$)?
 - (d) Compute subject A's estimated true standard score and its standard error ($\sigma_{\infty.1}$).
5. Show (a) that when the reliability coefficient is zero, the standard error of an obtained score equals the standard deviation of the test; and (b) that when the reliability coefficient is 1.00, the standard error of an obtained score equals zero.
6. A mathematics test has a reliability coefficient of .82, and a mechanical ability test has a reliability coefficient of .76. The r between the two tests is .52.
 - (a) What would the correlation between the two tests be if the mechanical ability test were a perfect measuring instrument (reliability coefficient of 1.00) [see p. 327]?
 - (b) What would the correlation between the two tests be if the mathematics test were a perfect measuring instrument?
 - (c) What would the correlation be if *both* tests were perfect measures?
7. A test of 40 items has a validity coefficient (i.e., correlation with a criterion) of .45, and a reliability coefficient of .75. If the test is lengthened to 120 items find
 - (a) the "new" validity coefficient;
 - (b) the "new" reliability coefficient;
 - (c) the maximum validity of which the given test is capable in terms of the present criterion.

8. An intelligence examination shows a correlation of .50 with first year scholarship. The reliability coefficient of the test is .85, and of the school grades (i.e., the criterion) is .65.
- What is the highest validity coefficient which one can hope to get with this test (i.e., when the test is a perfect measuring instrument)?
 - What is the maximum theoretical correlation (corrected correlation) between test and grades?
9. The reliability coefficient of Test Y in a large group is .92, the mean is 142, and the σ is 16. The reliability coefficient of Test X in the same group is .86, the mean is 54, and the σ is 10.
- What is the standard error of an obtained score in Test Y ?
 - What is the standard error of an obtained score in Test X ?
 - What is the standard error of an estimated true score in Test Y ?
 - What is the standard error of an estimated true score in Test X ?
 - In which test are the obtained scores more accurate estimates (p. 322)?

ANSWERS

- 6 times
 - $r_{11} = .75$ (doubling length); $r_{11} = .82$ (tripling length)
 - $PE_{.75} = .03$; $PE_{.82} = .03$
- .88
 - .90
 - In (a) $PE = .02$; in (b) $PE = .01$
- .89
 - 8.9
 - .64
- .49
 - .90
 - .52
- 9.14
 - .61; .19
 - 55.67; 2.69
 - .71; .28
- .54
 - .68
 - 4.5
 - 3.7
 - 4.2
 - 3.5
 - The first.
- .60
 - .57
 - .66

CHAPTER XII

THE INTERPRETATION OF THE COEFFICIENT OF CORRELATION

WHEN should a coefficient of correlation be called "high," when "medium," and when "low"? Does an r of .40 between two tests indicate "marked" or "low" relationship? How high should an r be in order to permit accurate prediction from one variable to another? Can an r of .50, say, be interpreted with respect to "overlap" of determining factors in the two variables correlated? Questions like these, all of which are concerned with the *significance* or *meaning* of the relationship expressed by a correlation coefficient, constantly arise in problems involving mental measurement, and their implications must be understood before we can effectively employ the correlational method.

The value of r as a measure of correspondence between two sets of measures may be profitably considered from two points of view.* In the first place, r 's are computed in order to determine whether there is *any* correlation (over and above chance) between two variables; and in the second place, r 's are computed in order to determine the *degree* or closeness of relationship when some association is known, or is assumed, to exist. The question, "Is there *any* correlation between brain weight and intelligence?", expresses the first objective. And the question, "How significant is the correlation between high school grades and first year performance in college?", expresses the second. The problem of when an obtained r denotes — or does not denote — significant relationship has already been considered in Chapter IX, page 280. Thus, we know that

* Barr, A. S., *The Coefficient of Correlation*, Journal of Educational Research, 1931, 23, pp 55-60.

when an r is four times its PE_r , it is significant, i e., expresses a relationship which is greater than chance. The present chapter will be concerned mainly with the second problem, namely, the meaning — with respect to *degree* of relationship — of an obtained coefficient. The questions at the beginning of the paragraph above all bear upon this topic. Several ways of answering such questions will now be considered.

I. VARIOUS INTERPRETATIONS OF THE COEFFICIENT OF CORRELATION

1. The Interpretation of r in Terms of Verbal Description

In the field of mental measurement it is customary to describe the correlation between two tests in a general way as being high, marked or substantial, low or negligible. While the descriptive label applied will vary somewhat in meaning with the author using it, there is fairly good agreement among workers with psychological and educational tests that an

r from .00 to $\pm .20$ denotes indifferent or negligible relationship;
 r from $\pm .20$ to $\pm .40$ denotes low correlation; present but slight,
 r from $\pm .40$ to $\pm .70$ denotes substantial or marked relationship;
 r from $\pm .70$ to ± 1.00 denotes high to very high relation.

This classification is broad and somewhat tentative, and can be accepted as a general guide only in the light of various qualifications. A coefficient of correlation should always be evaluated with regard to (1) the nature of the material dealt with; (2) PE_r ; (3) the size and variability of the group (p. 303); (4) the reliability coefficients of the tests (p. 335); and (5) the purpose for which the r was computed. To consider, first, the matter of material, an r of .30 between height and intelligence, or between head measurements and mechanical ability if significant — more than four times PE — would be regarded as marked rather than low, since experience has shown * that correlations between physical and mental func-

* Garrett, H. E., and Schneek, M. R., *Psychological Tests, Methods and Results*, 1933, Part I, Chapter 1.

tions are usually much lower — around zero. On the other hand, the correlation must be .70 or more between general intelligence measures and school grades or between achievement in English and history to be considered high, since r 's in this field usually run from .40 to .60. Resemblance, with respect to physical and mental traits, of siblings and of parents and offspring is expressed by r 's of from .35 to .55; and, accordingly, an r of .60 here would be high.* By contrast, the correlations among standard intelligence tests are ordinarily so much higher than .60, that a correlation coefficient between two such measures must be .80 to .90 to be regarded as high. In the field of vocational testing, the r 's between test batteries and measures of aptitude represented by various criteria are rarely above .50 †; and r 's above this figure would be considered surprisingly high.

Correlation coefficients must also be evaluated with due regard to the reliabilities of the two tests concerned. Because of chance errors, an obtained r is always less than its "corrected" value (p. 335) and hence, in a sense, is a minimum measure of the relationship present. The effect upon an r of the size and variability of the group has been discussed elsewhere (p. 303), and a formula for evaluating such effect provided. Finally, the *purpose* for which an r was computed is important. Thus, the r which is to be employed as a means of predicting the scores of *individuals* from one test to another (by means of the regression equation) must be much higher than the r the purpose of which is simply to provide a measure of the relationship between two tests of achievement in the group as a whole.

* Jones, H. E., *A First Study of Parent-Child Resemblance in Intelligence*, 27th Yearbook of the N S S E, 1928, Part I, pp. 61-72.

Thorndike, E. L., *The Resemblance of Siblings in Intelligence*, 27th Yearbook of the N S S E, 1928, Part I, pp. 41-53.

† Hull, C. L., *Aptitude Testing*, 1928, pp. 273-275.

2. The Interpretation of r in Terms of $\sigma_{(\text{est. } Y)}$ and the Coefficient of Alienation

Perhaps the most practical way of evaluating the effectiveness of a coefficient of correlation as a means of predicting individual scores is through the standard error of estimate, $\sigma_{(\text{est. } Y)}$. We have already found (p. 300) that $\sigma_{(\text{est. } Y)}$ — which equals $\sigma_y \sqrt{1 - r^2}$ — enables us to tell how accurately we can estimate, by means of the regression equation, an individual's score in Test Y when we know his score in Test X . The size of $\sigma_{(\text{est. } Y)}$ depends directly upon σ_y and upon the correlation between the two tests. When $r = 1.00$, $\sigma_{(\text{est. } Y)} = .00$, which means that we can predict a person's score in Y , knowing his score in X , with complete accuracy — no error. On the other hand, when $r = .00$, $\sigma_{(\text{est. } Y)} = \sigma_y$, which means that we can only be certain that the predicted score lies *somewhere* within the limits of the Y -distribution, i.e., within the limits Mean Score $\pm 3\sigma_y$. In other words, when $r = .00$ our estimate of a Y -score is not aided at all by a knowledge of the distribution of X . As r decreases from 1.00 to .00, the standard error of estimate increases so rapidly that predictions from the regression equation range all the way from certainty to what is virtually a "guess." * The significance of an r , with respect to predictive value, therefore, may be accurately gauged by the size of $\sigma_{(\text{est. } Y)}$.

The following problem will serve as an illustration. Suppose that the correlation between two tests Y and X is .60, and that $\sigma_y = 5.00$. Then $\sigma_{(\text{est. } Y)}$ is $5 \times \sqrt{1 - .60^2}$ or 4.00, which is 20 percent less than 5.00, the $\sigma_{(\text{est. } Y)}$ when $r = .00$ — i.e., when $\sigma_{(\text{est. } Y)}$ has minimum predictive value. The amount of reduction in $\sigma_{(\text{est. } Y)}$ as r varies from .00 to 1.00 is given by the expression $\sqrt{1 - r^2}$, and hence it is possible from $\sqrt{1 - r^2}$ alone to estimate the predictive value of an r . The expression

* The term "guess" as here used does not imply an estimate which is based upon no information whatsoever — a shot in the dark, so to speak. When $r = .00$, the most probable Y -score predicted for every individual in the X -distribution is \bar{M}_Y , and $\sigma_{(\text{est. } Y)} = \sigma_y$. Hence, our Y -estimates are "guesses" in the sense that they *may* lie anywhere in the Y -distribution — but not anywhere at all!

$\sqrt{1 - r^2}$ has been called the *coefficient of alienation* * and is usually denoted by the letter k . The coefficient of alienation may be thought of as measuring the *absence* of relationship between two variables X and Y in the same sense in which r measures the *presence* of relationship. When $k = 1.00$, $r = .00$, and when $k = .00$, $r = 1.00$, so that the *larger* the coefficient of alienation the *smaller* the degree of relation, and the less precise the prediction from X to Y . In order to show how the estimate improves as r increases, the k 's for certain values of r from .00 to 1.00 are tabulated in Table 42.

TABLE 42
GIVING COEFFICIENTS OF ALIENATION k FOR VALUES OF r
FROM .00 TO 1 00

r	$k = \sqrt{1 - r^2}$	r	$k = \sqrt{1 - r^2}$
0000	1.0000	.8000	.6000
10	.9950	.8660	.5000
20	.9798	.90	.4359
30	.9539	.95	.3122
40	.9165	.98	.1990
.50	.8660	.99	.1411
60	.8000	1.00	0000
70	.7141		
(.7071)	.7071		

It will be noted that r must be .866 before k lies *half way* between 1.00 and .00 — before the standard error of estimate is reduced one-half. For r 's of .80 or less, the coefficients of alienation are clearly so large that predictions of individual scores based upon the regression equation are little better than a "guess." Even when $r = .99$, the standard error of estimate is still $\frac{1}{4}$ as large as when $r = .00$. In order to estimate accurately individual scores in Y from a knowledge of X , therefore, the correlation between Y and X should be .90 or more.

The coefficient E given by the formula below is often useful in providing a quick estimate of the predictive efficiency of an obtained r . E , which is called the coefficient of "forecasting

* Kelley, T. L., *Principles Underlying the Classification of Men*, Journal of Applied Psychology, 1919, 3, pp. 50-67.

efficiency" * or the coefficient of dependability, is derived from k as follows:

$$E = 1 - \sqrt{1 - r^2} \quad (89)$$

or

$$E = 1 - k$$

(coefficient of "forecasting efficiency" or coefficient of dependability) †

To illustrate the application of E , suppose that the correlation of a test (or of a test battery) with some criterion of performance is .50. From formula (89) $E = 1 - .87$ or .13; and the test's efficiency in predicting criterion scores may be said to be 13 percent. When $r = .90$, $E = .56$ and the test is 56 percent efficient; when $r = .98$, $E = .80$ and the test is 80 percent efficient, and so on. Obviously, the correlation must be above .87 for the test's forecasting efficiency to be greater than 50 percent.

It will be noted that E gives essentially the same information as $\sigma_{(\text{est } Y)}$ or k . Thus, if $r = .50$, $k = .87$ and $\sigma_{(\text{est } Y)}$ is 87 percent of σ_y , which is its value when $r = .00$. Accordingly, an $r = .50$ reduces the $\sigma_{(\text{est } Y)}$ by 13 percent.

3. The Interpretation of r in Terms of the Standard Error of an Obtained Score

We saw on page 321 that the standard error of an individual score enables us to estimate the probable divergence of a score obtained on a test from its corresponding true score. Since $\sigma_{1\infty} = \sigma_1 \sqrt{1 - r_{11}}$ (p. 320), the probable divergence of an obtained score from its theoretically true value depends upon σ_1 and upon r_{11} , the reliability coefficient of the test. Since σ_1 is constant for the given test, the value of $\sigma_{1\infty}$, as a measure of the closeness of correspondence of obtained and true scores, may be determined by the size of r_{11} . When $r_{11} = 1.00$, for

* Hull, C. L., *Aptitude Testing*, 1928, Chapter 8, pp. 268-276.

† See also Conrad, H. S., and Martin, G. B., *The Index of Forecasting Efficiency, for the Case of a "True" Criterion*, Journal of Experimental Education, 1935, 4, pp. 231-244.

example, $\sigma_{1\infty} = .00$, and every obtained score equals its true score exactly. When $r_{11} = .00$, on the other hand, $\sigma_{1\infty} = \sigma_1$ (the σ of the distribution) and we can only be sure that the true score, corresponding to the given obtained score, lies somewhere within the limits of the distribution, within the limits Mean Score $\pm 3\sigma_1$. In other words, when $r_{11} = .00$, the probable divergence of an obtained score from its theoretically true value is as great as it would be had we simply "guessed" that the true score lay somewhere in the distribution.

To illustrate with an example, suppose that the reliability coefficient of a given test, r_{11} , equals .80 and that σ_1 equals 10.00. Then $\sigma_{1\infty} = 10\sqrt{1 - .80}$ or 4.47 and since σ_1 is 10.00 when $r_{11} = .00$, it is evident that a reliability coefficient of .80 serves to reduce $\sigma_{1\infty}$ by 55 percent, or to about 45 percent of what it would be if the self-correlation were zero. The reduction in $\sigma_{1\infty}$ as r_{11} varies from .00 to 1.00 is given by the expression, $\sqrt{1 - r_{11}}$. Hence, this factor may be used to measure the effectiveness of an obtained reliability coefficient in reducing $\sigma_{1\infty}$, just as k is used to measure the effectiveness of r_{xy} in reducing $\sigma_{(\text{est } Y)}$. In Table 43 the values of $\sqrt{1 - r_{11}}$ have been tabulated for selected r_{11} 's from .00 to 1.00. From the table it is clear that the self-correlation of a test must be at least .75 before $\sqrt{1 - r_{11}}$ is half way between 1.00 and .00 — before the standard error lies half way between a standard error which provides a "guess" estimate, and one which provides a 100%

TABLE 43
GIVING VALUES OF $\sqrt{1 - r_{11}}$ FOR VALUES OF r_{11}
FROM .00 TO 1.00

r_{11}	$\sqrt{1 - r_{11}}$	r_{11}	$\sqrt{1 - r_{11}}$
0000	1.0000	.8000	.4472
.10	.9487	.90	.3162
.20	.8944	.95	.2236
.30	.8367	.98	.1414
.40	.7746	.99	.1000
.50	.7071	1.00	.0000
.60	.6325		
.70	.5477		
.75	.5000		

accurate estimate. When $r_{11} = .98$, the chances are still 68 in 100 that a given obtained score will diverge from its true counterpart by as much as .1414 times the σ_1 of the test. Since even with such extremely high reliability coefficients, gaps still remain between obtained and true scores, it is obvious that reliability coefficients below .90 are of little value if we require the accurate determination of individual scores.

4. The Interpretation of r in Terms of Common Factors *

It is often enlightening to think of a coefficient of correlation as an index which expresses the degree of "communality" among the elements or factors in the two tests which are being correlated. Again, r may be thought of as indicating the extent to which the factors determining the score in one test "overlap" the determining factors of another test. Let us suppose that performance in X depends upon the presence or absence of $a + c$ independent, elemental factors; and that performance in Y depends upon the presence or absence of $b + c$ independent, elemental factors. The a factors determine X measures alone, the b factors Y measures alone, and the c factors are common to both X and Y . Moreover, let us suppose that all factors, a , b , and c , are governed solely by the laws of chance, so that each factor is as likely to be present as absent, in the same way that a coin when tossed is as likely to fall heads as tails.

Now if we let n_a = the total number of a factors, n_b = the total number of b factors, and n_c = the total number of c factors, the correlation between X and Y is given by the formula†:

$$r = \frac{n_c}{\sqrt{(n_a + n_c)(n_b + n_c)}} \quad (90)$$

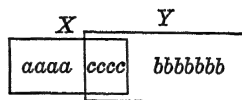
That is, the coefficient of correlation equals the number of common factors in X and Y , divided by the geometrical mean of the

* For further discussion of this topic, see: Tryon, R. C., *The Interpretation of the Correlation Coefficient*, Psychological Review, 1929, 36, pp. 419-445.

Dunlap, J. W., and Cureton, E. E., *On the Analysis of Causation*, Journal of Educational Psychology, 1930, 21, pp. 657-680.

† Brown, Wm., and Thomson, G. H., *The Essentials of Mental Measurement*, 1925, p. 141.

total number of factors in X and Y . This situation is shown graphically in Figure 50 in which X is determined by 8 factors,



$$r = \frac{4}{\sqrt{8 \times 11}} = .426$$

FIG. 50.

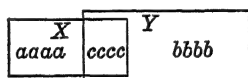
4 a 's and 4 c 's, and Y by 11 factors, 7 b 's and 4 c 's. The correlation by formula (90) is

$$r = \frac{4}{\sqrt{(4+4)(7+4)}} \text{ or } \frac{4}{\sqrt{8 \times 11}} = .426$$

If the number of elementary factors determining the score in X equals exactly the number determining the score in Y , so that $n_a = n_b$, formula (90) becomes

$$r = \frac{n_c}{n_a + n_c} \quad (91)$$

and the coefficient of correlation is now simply the decimal fraction which indicates what *proportion* of the "causes" influencing performance in X and Y are common to both. This situation



$$r = \frac{4}{8} = .50$$

FIG. 51.

is illustrated in Figure 51. Since X is determined by 8 factors, 4 a 's and 4 c 's, and Y by 8 factors, 4 b 's and 4 c 's, the correlation by formula (91) is $\frac{4}{8}$ or .50.

Now let us assume, lastly, that Y is *completely* determined by n_c elements, and that X is determined by these same elements

plus n_a elements in addition ($n_b = 0$). Formula (91) then becomes

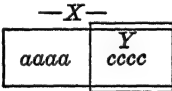
$$r = \frac{n_c}{\sqrt{n_c(n_a + n_c)}} \quad (92)$$

and the coefficient of correlation equals the number of common elements in X and Y divided by the geometrical mean of the total number of factors in X and in Y . Figure 52 illustrates this situation graphically. Y is determined by 4 c 's, and X by these factors plus 4 a 's in addition; and the correlation is

$\frac{4}{\sqrt{4 \times 8}}$ or .707. If we square the r obtained from formula (92), we have

$$r^2 = \frac{n_c}{n_a + n_c} \quad (93)$$

that is, the square of the coefficient gives the extent to which the elements in Y overlap those of X , or the proportion of elements in X which are also present in Y . In Figure 52 note that



$$r = \frac{4}{\sqrt{4 \times 8}} = .707$$

FIG. 52.

Y overlaps 50% of X and that r^2 , or $(.707)^2$, is .50, as it should be. Moreover, since the coefficient of alienation equals .707 when $r = .707$ (see Table 42), it follows that an r of .707 (and not of .50) should be taken as *half* of a perfect correlation.* On the same assumptions, an overlapping of 33⅓% common elements ($r^2 = .333$) will give a correlation of .578, which is ⅓ of a perfect correlation; and an overlapping of 25% common elements ($r^2 = .25$) gives an $r = .50$, which is ¼ of a perfect correlation. By analogy, an r of .30 or less implies so slight a

* Hull, C. L., *The Joint Yield from Teams of Tests*, Journal of Educational Psychology, 1923, 14, pp. 396-406.

Woodworth, R. S., *Combining the Results of Several Tests. A Study in Statistical Method*, Psychological Review, 1912, 19, p. 113

degree of overlapping that there can be a very small percentage of common elements.

Looked upon as a measure of the percentage of common factors or elements in two variables, the coefficient of correlation may be studied most advantageously, perhaps, when calculated between series formed by tossing coins or throwing dice. In such data the degree of "overlapping" is definitely known. To illustrate, consider the correlation table in Figure 53 in

		HEADS IN FIRST TOSS														
HEADS IN SECOND TOSS		0	1	2	3	4	5	6	7	8	9	10	11	12	Total	
	12															
	11								1						1	
	10					1		2		2	3	1	1		10	
	9						2	9	13	4	3				31	
	8				1	5	9	10	18	14	4	2			63	
	7			1	2	5	14	24	28	10	7	4			95	
	6			1	3	9	18	27	29	16	3	2	1		109	
	5				4	11	23	21	15	9					83	
	4				3	6	9	21	14	10	5	1			69	
	3				3	3	8	4	4	4					26	
	2				3	1	5	1	1						11	
	1						1	1			.				2	
	0															
	Total				11	20	54	93	112	118	60	21	9	2		500

X							Y						
a	a	a	a	a	a	a	c	c	c	c	c	b	b
b	b	b	b	b	b	b	b	b	b	b	b	b	b

$$\begin{aligned}
 n_c &= 5 \\
 n_a &= n_b = 7 \\
 r &= \frac{n_c}{n_a + n_c} = \frac{5}{12} = .417
 \end{aligned}
 \tag{91}$$

By calculation (product-moment)

$$r = .432$$

FIG. 53. Showing the Number of Heads in 500 Successive Throws of 12 Pennies in which 7 Pennies were Tossed in the Second Throw and 5 Remained as they Fell in the First Throw of all 12 Together.*

* From Pearl, R., *Medical Biometry and Statistics*, 1930, p. 370 (after Darbishire).

FIRST THROW OF 5 DICE (X)

SECOND THROW OF 10 DICE (Y)	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Total
45												1	1	1		3
44									1							1
43								2				1		2		5
42								1	1	1		1	1		1	6
41										1			1			2
40									2	3			1			6
39									2	2	1			1		6
38							1		1	1	2	1				6
37			1					1	2	2	1	1			1	9
36					1	1	2	2	1	1						8
35					1		1	2	1	1	2		1			9
34			1	1			2	1	1	2		1				9
33	1			1					1							3
32						1	2	1				1				5
31			1	1		1			1	1						5
30					2		1	1								4
29	1							1								2
28			2		1	1	1									5
27			2				1									3
26	1															1
25		1			1											2
Total	3	1	7	3	6	4	11	12	14	15	6	6	5	4	3	100

By calculation (product-moment)

$$r = .694$$

$$n_c = 5$$

$$n_a = 5$$

—X—									
(5)					(5)				
a	a	a	a	a	c	c	c	c	c
—Y—									

$$r = \frac{n_c}{\sqrt{n_c(n_a + n_c)}} = \frac{5}{\sqrt{5 \times 10}} = .707 \quad (92)$$

FIG. 54. Showing the Results of 100 Successive Throws of Dice in First Throw of which (X) 5 Dice were Thrown, Counted, and Left Down; and in Each Second Throw of which (Y) 5 Additional Dice were Thrown and Counted Together with the 5 Left Down (10 in all).

which is shown the relation between two series of 500 successive throws of 12 pennies made in the following way: first, all 12 pennies were tossed, and the number of heads recorded and noted in the X-column; then 5 coins were left lying and the

remaining 7 were tossed again and the number of heads in all 12 recorded and noted in column *Y*, opposite the *X*-entry. By this scheme, 5 coins (factors) contribute to each pair of tosses and hence, according to formula (91) the correlation should be $\frac{5}{12}$ or .417. By the product-moment formula the actual correlation between the two series is .432, which indicates a very close correspondence between actual and theoretical results. The situation existing in each pair of *X* and *Y* tosses is shown in Figure 53. If 4 coins had been left lying, the *r* would have been $\frac{4}{12}$ or .333; if 6 had been left lying, *r* would have been $\frac{6}{12}$ or .50 and so on. A number of diagrams of the sort shown in which the number of common factors (i.e., coins left lying) varies from 0 to 12, and *r* from 0 to 1.00 may be found in Pearl's *Medical Biometry and Statistics*, pages 367-374.

Now suppose that we calculate the correlation between two series of dice throws made according to the following scheme * 5 dice were thrown, and the total read and recorded in the *X* column; then 5 additional dice were thrown and the total (the 5 left and the 5 just thrown) were read and recorded in the *Y*-column. This process was continued until 100 throws had been made, each *Y*-throw (of 10 dice) "overlapped" to the extent of 50% by its corresponding *X*-throw (of 5 dice). Since one-half of the elements in *Y* are contained in *X*, the correlation between *Y* and *X* should, by formula (92), be $\frac{5}{\sqrt{5 \times 10}}$ or .70

(see Fig. 54). Actually, the correlation by the product-moment method was .694, which indicates, again, a very close correspondence between actual and theoretical results. The square of this *r* gives us approximately .50 as the percentage of elements in *Y* which are also in *X*; that is, we have one-half of a perfect correlation (p. 350).

While formulas (90 to 93) provide an interesting way of interpreting a coefficient of correlation under certain arbitrary conditions, it would be a mistake to assume that by simply squaring any correlation coefficient we can determine forthwith the per

* These throws were made by the writer.

centage of factors in the one test ability which are also operative in the second test ability. Scores upon psychological and educational tests are very probably conditioned by the combined action of many factors interwoven in a relatively complex manner. The correlation between two tests may be the same, whether the factors are independent and additive, or dependent and related (provided they are numerous), so that one cannot reason from identity of result to identity of "causal" factors.

5. The Interpretation of r in Terms of the Coefficient of Determination

The interpretation of r in terms of "overlapping" factors in the two tests being correlated may be generalized somewhat through an analysis of the *variance* (σ^2) of the dependent variable — usually the Y test. In studying the variability among individuals upon a given test, the variance of the test scores is often a more useful measure of "spread" than is the standard deviation.* The total variance of Test Y may be thought of as depending upon the variability in Test X and upon other factors not involved in X . The object of analyzing the variance of Test Y , therefore, is to determine from the correlation between Y and X what part of Test Y 's variance is associated with, or dependent upon, the variance of Test X , and what part is determined by factors not in Test X .

If we have given the correlation between the two Tests Y and X , σ^2_y gives a measure of the total variance of the Y -scores; and $\sigma^2_{(\text{est } Y)}$, which equals $\sigma^2_y(1 - r^2_{xy})$, gives a measure of the variance *left* in Test Y when that part of the variance produced by Test X is *ruled out* or *held constant*.† To illustrate, if we have the correlation between height and weight in a group of school children, $\sigma^2_{(\text{ht})}$ will be reduced to $\sigma^2_{(\text{est. ht})}$, when the variance in weight is zero — when *all* of the children have the same weight. If $\sigma^2_{(\text{est. } Y)}$ is subtracted from σ^2_y there re-

* Ezekiel, M., *Methods of Correlation Analysis*, 1930, p. 120 and pp. 375 ff

† See Chapter XIV for further discussion of this topic.

mains that part of the variance of Test Y which is associated with Test X ; and if this value is divided by σ_y^2 , we obtain that fraction or percentage of the variance of Test Y *attributable to or associated with* Test X . Carrying out the operations described, we have

$$\frac{\sigma_y^2 - \sigma_{(est\ Y)}^2}{\sigma_y^2} = \frac{\sigma_y^2 - \sigma_y^2 + \sigma_y^2 r^2_{xy}}{\sigma_y^2} = r^2_{xy}$$

from which it is clear that r^2_{xy} gives the *percentage* of the variance of Test Y which is associated with Test X . When used for this purpose, r^2 is often called the *coefficient of determination*. If the correlation between Tests Y and X is .707, r^2 is .50. Hence, an r of .707 means that 50% of the variance of Test Y is associated with the variability in Test X . Since $r^2 + k^2 = 1.00$, the proportion of the variance in Test Y which is *not* associated with Test X is given by k^2 . In the present case, since r^2 is .50, k^2 is also .50.

The interpretation of r^2 as giving the percentage of the variance of the dependent variable which is associated with the independent variable is a more general result than is the interpretation of r^2 as giving the percentage of elements in one test which are also in the other test. In taking the second view, we must restrict our conclusion to those situations in which *all* of the factors determining performance in Test Y , say, are also found in Test X (see Fig. 54, p 352). The coefficient of determination, however, will always tell us what part of the variance of Test Y is determined by Test X —but no information, of course, is given as to the character of the association. Inspection of the squares of small coefficients of correlation emphasizes again the slight degree of association, in terms of related changes in test variability, indicated by low r 's. An r of .10, for example, or .20, or .30, between Tests X and Y , indicates that only 1%, 4%, and 9%, respectively, of the variance of Y is associated with X . On the other hand, when r is .95, about 90% ($r^2 = .90$) of the variance of Test Y is associated with Test X , only 10% being totally unrelated. Valuable insight into the part played by one or more variables

in determining the total variance of a criterion may be obtained through the coefficient of determination (p. 448).

6. Summary

It will be helpful to draw together, in the form of a general summary, the main points brought out in this chapter.

- (1) Whether an obtained r is to be regarded as "high," "medium" or "low" will depend upon the material which is being studied, the reliability coefficients of the two tests, the size of the group and its variability, and the purpose for which the r is being computed.
- (2) The significance of an r as a measure simply of the *presence* of relationship may be evaluated in terms of PE_r .
- (3) The accuracy with which an r enables us to predict *individual* scores (through the regression equation) in Test Y from given scores in Test X may be determined from $\sigma_{(\text{est } Y)}$, from E , or from the *coefficient of alienation*.
- (4) The significance of r_{II} , the reliability coefficient, as a measure of the extent to which obtained scores probably diverge from their true counterparts is given by $\sqrt{1 - r_{II}}$.
- (5) Under certain assumptions as to independence of "causes," r may be interpreted in terms of overlapping elements or factors common to the two tests correlated (see formulas 90 to 93)
- (6) The *coefficient of determination* provides a way of telling what percentage of the total variance (σ^2) of Test Y is associated with Test X ; and what percentage is independent of Test X . This method of analysis may be extended to problems employing partial and multiple correlation.*

* Wright, Sewall, *Correlation and Causation*, Journal of Agricultural Research, 1921, 20, pp. 557-585

Burks, B. S., *The Relative Influence of Nature and Nurture upon Mental Development, a Comparative Study of Foster Parent-Foster Child Resemblance, and True Parent-True Child Resemblance*, 27th Yearbook of the N.S.S.E., 1928, Part I, pp. 219-316.

Heilman, J. D., *The Relative Influence upon Educational Achievement of Some Hereditary and Environmental Factors*, 27th Yearbook of the N.S.S.E., 1928, Part II, pp. 35-66.

PROBLEMS

1. Basing your answer upon your experience and general knowledge of psychology, decide whether the correlation between the following pairs of variables is most probably (1) positive or negative; (2) high, medium, or low.
 - (a) Intelligence of husbands and wives.
 - (b) Brain weight and intelligence.
 - (c) High school grades in history and physics.
 - (d) Age and radicalism.
 - (e) Extroversion and college grades.
2. How much more will an r of .80 reduce a given $\sigma_{(est.)}$ than an r of .40? An r of .90 than an r of .40?
3. (a) Determine k and E for the following r 's. .35, — .50; .70; .95.
 (b) What is the "forecasting efficiency" of an r of .45? an r of .99?
4. The reliability coefficient of a test has been raised from .60 to .80 by the process of lengthening the test. By how much is the σ_{100} reduced?
5. One hundred throws of 10 pennies were made in the following way: All coins were thrown, and the number of heads recorded; six coins were allowed to lie as they fell, and the remaining four thrown again, all ten being then read a second time. What is the "theoretical" r between the first and second series of throws?
6. Performance in Test Y is determined by 35 elemental factors; and performance in Test X by 42 elemental factors. If there are ten factors common to the two tests, what is r_{xy} ?
7. The correlation of a criterion with a test battery is .75. What percent of the variance of the criterion is associated with variability in the battery? What percent is independent of the battery?
8. The σ of a certain test is 10. What must r_{11} be in order that the odds will be roughly 2 : 1 that an obtained score will not diverge from its true value by more than 3 points?
9. The σ of Test Y is 15. What must the correlation between Tests X and Y be in order that the odds may be even (50 : 50)?

that a score' on Test Y predicted from a score on Test X will not diverge from its actual value by more than 5 points?

10. Interpret a coefficient of .60 in at least four ways.

ANSWERS

2. Five times as much; seven times as much.

3. (a)	r	k	E
	.35	.94	.06
	— .50	.87	.13
	.70	.71	.29
	.95	.31	.69

(b) 11%; 86%

4. About 29%

5. $r = .60$

6. $r = .26$

7. 56%; 44%

8. $r_{11} = .91$

9. $r_{xy} = .87$

CHAPTER XIII

FURTHER METHODS OF CORRELATION

IN Chapters IX and X, we described the linear, or product-moment correlation method, and showed how, by means of r and the regression equations, one can "predict" values of the one variable from a knowledge of the other. The linear correlation coefficient is useful in psychology and education as a measure, primarily, of the relationship between test scores and other measures of performance. Test scores (as we have seen) represent a series of determinations of a continuous variable taken along a numerical scale. Many situations arise, however, in which the investigator does not have scores and must work with data in which differences in merit or capacity can be expressed only by ranks (e.g., in orders of merit); or by classifying an individual into one of several descriptive categories. This is especially true in vocational and applied psychology and in the field of personality and character measurement. Again, there are problems in which the relationship among the measurements made is *non-linear*, and hence cannot be described by the product-moment r . In cases of this sort other methods of determining correlation must be employed; and the purpose of this chapter is to develop some of the more useful of these techniques.

I. METHODS OF MEASURING CORRELATION WHICH TAKE ACCOUNT ONLY OF RELATIVE POSITION OR RANK

It will be readily seen that differences among individuals in many traits can be expressed by ranking the subjects in one-two-three order, when such differences cannot be measured directly. For example, persons may be ranked in order of merit

for honesty, athletic ability, salesmanship, or social adjustment. In like manner, various products or specimens such as advertisements, color combinations, handwriting specimens, compositions, jokes, and pictures may be ranked in order for esthetic quality, beauty, humor, or some other characteristic. In computing the correlation between two series of ranks, special methods which take account of relative position have been devised. These methods may also be applied to *scores* which have been arranged in order of merit. When we have only a few scores (less than 25 for example), although these represent quantitative determinations on a linear scale, it is often advisable to rank them in order of merit and compute the correlation by a rank method instead of by the longer and more laborious product-moment method. Coefficients of correlation calculated from a few cases are never very reliable at best, and are often of value chiefly in suggesting the possible existence of relationship, or in a preliminary survey. In such cases rank methods are recommended. They will probably give as good a result as that obtained by a more refined technique, and are much easier to apply.

In this section we shall consider two methods of computing the correlation when the data to be correlated are arranged in orders of merit. The first is the *method of rank-differences*; and the second is the *method of gains* or the Spearman "foot-rule."

1. The Method of Rank-Differences

The method of rank-differences is illustrated in Table 44. The problem is to find the relationship between the length of service and the selling-efficiency of 12 salesmen. The names of the men (A, B, C, etc.) are listed in column 1 of the table, and in column 2, opposite the name of each man, is given the number of years he has been in the service of the company. In column 3, the men are ranked in order of merit in accordance with their length of service. For example G, who has been longest with the company, is ranked 1; C, whose length of service is next longest, is ranked 2; and so on down the list.

TABLE 44
TO ILLUSTRATE THE RANK-DIFFERENCE METHOD OF
MEASURING CORRELATION

(1)	(2)	(3)	(4)	(5)	(6)
Salesmen	Years of Service	Order of Merit (Service)	Order of Merit (Efficiency)	Difference between Ranks (<i>D</i>)	Difference Squared (<i>D</i> ²)
A	5	7.5	6	1.5	2.25
B	2	11.5	12	.5	.25
C	10	2	1	1.0	1.00
D	8	4	9	5.0	25.00
E	6	6	8	2.0	4.00
F	4	9	5	4.0	16.00
G	12	1	2	1.0	1.00
H	2	11.5	10	1.5	2.25
I	7	5	3	2.0	4.00
J	5	7.5	7	.5	.25
K	9	3	4	1.0	1.00
L	3	10	11	1.0	1.00
<i>N</i> = 12					58.00

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)} = 1 - \frac{6 \times 58}{12(143)} = .80 \quad (94)$$

From Table 45, $r = .81$.

$$PE\rho = \frac{.7063(1 - \rho^2)}{\sqrt{N}} = .07 \quad (95)$$

Note that both A and J have the same period of service, and that each is ranked 7.5. Instead of ranking the first man 7 and the second man 8, or both 7 or both 8, we compromise by ranking both 7.5 and F, who follows, 9 (see p. 190).

In column 4 the men have been ranked by the sales manager in order of merit for efficiency as salesmen. C, the most efficient man, is ranked 1; and B, the least efficient, is ranked 12. In column 5 the difference between each man's efficiency rank and his years-of-service rank (designated *D*) is entered; and in the last column each of these *D*'s has been squared.* The correlation between the two orders of merit may now be computed by substituting for $\sum D^2$ and *N* in the formula

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)} \quad (94)$$

(rank correlation coefficient, ρ)

* Since each *D* is squared in column 6, no account need be taken of + and - signs in column 5.

in which D represents the difference in the rank of an individual in the two series; $\sum D^2$ is the sum of the squares of all such differences; and N is the number of cases. The symbol ρ (read as rho) is the rank order coefficient of correlation; ρ may be transmuted into a product-moment r by means of Table 45.*

TABLE 45

A TABLE TO INFER THE VALUE OF r FROM A GIVEN VALUE OF ρ

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

ρ	r	ρ	r	ρ	r	ρ	r
.01	.0105	.26	.2714	.51	.5277	.76	.7750
.02	.0209	.27	.2818	.52	.5378	.77	.7847
.03	.0314	.28	.2922	.53	.5479	.78	.7943
.04	.0419	.29	.3025	.54	.5580	.79	.8039
.05	.0524	.30	.3129	.55	.5680	.80	.8135
.06	.0628	.31	.3232	.56	.5781	.81	.8230
.07	.0733	.32	.3335	.57	.5881	.82	.8325
.08	.0838	.33	.3439	.58	.5981	.83	.8421
.09	.0942	.34	.3542	.59	.6081	.84	.8516
.10	.1047	.35	.3645	.60	.6180	.85	.8610
.11	.1151	.36	.3748	.61	.6280	.86	.8705
.12	.1256	.37	.3850	.62	.6379	.87	.8799 ✓
.13	.1360	.38	.3955	.63	.6478	.88	.8893
.14	.1465	.39	.4056	.64	.6577	.89	.8986
.15	.1569	.40	.4158	.65	.6676	.90	.9080
.16	.1674	.41	.4261	.66	.6775	.91	.9173
.17	.1778	.42	.4363	.67	.6873	.92	.9269
.18	.1882	.43	.4465	.68	.6971	.93	.9359
.19	.1986	.44	.4567	.69	.7069	.94	.9451
.20	.2091	.45	.4669	.70	.7167	.95	.9543
.21	.2195	.46	.4771	.71	.7265	.96	.9635
.22	.2299	.47	.4872	.72	.7363	.97	.9727
.23	.2403	.48	.4973	.73	.7460	.98	.9818
.24	.2507	.49	.5075	.74	.7557	.99	.9909
.25	.2611	.50	.5176	.75	.7654	1.00	1.0000

Substituting 58 for the $\sum D^2$ and 12 for N in formula (94), we obtain a ρ of .80; and from Table 45 a ρ of .80 is found to be equivalent to an r of .81. The PE of ρ is about 5% larger than the PE of the product-moment r . The formula is

$$PE_{\rho} = \frac{.7063(1 - \rho^2)}{\sqrt{N}} \quad (95)$$

(probable error of ρ , the rank-order correlation coefficient)

* The difference between ρ and its equivalent r is so small, that for most purposes ρ may be taken as equal directly to r .

and, since in the present example ρ is equal to .80, $PE_{\rho} = .07$. The coefficient of correlation, though based on only 12 cases, is conventionally reliable. Whenever N is less than 30, however, the PE of a correlation coefficient is probably much larger than the value given by the formula. For this reason, a correlation coefficient calculated from less than 30 cases should always be interpreted with caution. In the present example there is good evidence of close correspondence between rankings for efficiency and number of years employed; but the sample is not large enough to make our finding very general.

2. The Method of Gains or the Spearman Foot-Rule

A second method of computing correlation when the data are ranked in order of merit is the *method of gains* or the Spearman "foot-rule." * Table 46 illustrates the use of the foot-rule as applied to the data in Table 44. The first four columns are the

TABLE 46

TO ILLUSTRATE THE FOOT-RULE METHOD OF FINDING CORRELATION

(1) Salesmen	(2) Years of Service	(3) Order of Merit (Service)	(4) Order of Merit (Efficiency)	(5) G (Gains) (4 over 3)	(6) G (Gains) (3 over 4)
A	5	7.5	6	1.5	
B	2	11.5	12		.5
C	10	2	1	1.0	
D	8	4	9		5.0
E	6	6	8		2.0
F	4	9	5	4.0	
G	12	1	2		1.0
H	2	11.5	10	1.5	
I	7	5	3	2.0	
J	5	7.5	7	.5	
K	9	3	4		1.0
L	3	10	11		1.0
				<u>10.5</u>	<u>10.5</u>

$$R = 1 - \frac{6\Sigma G}{N^2 - 1} = 1 - \frac{6 \times 10.5}{143} = .56$$

$$r \text{ (Table 47)} = .79$$

* Spearman, C., "Footrule" for Measuring Correlation, British Journal of Psychology, 1906, 2, pp. 89-108.

same as in the rank-difference method; as before in columns 3 and 4 each series is arranged in an order of merit. But the methods differ from here on. The entries in column 5 which is headed *G* (Gains) are found by taking the *plus* differences or the gains in rank of the 12 men in efficiency ranking as compared with their service ranking. For example, A who ranks 7.5 from the top in "service" and 6 from the top in "efficiency" has a *plus* gain in rank of 1.5 from the second ranking to the first.* C, F, H, I, and J likewise register plus differences or gains in their "efficiency" rankings as compared with their "service" rankings. The total of the *G* column is 10.5. If we compute gains in service over efficiency, instead of in efficiency over service, the same *G* will be obtained, as shown in column 6. It makes no difference, therefore, whether we figure gains in terms of the first series over the second, or the other way around, the second over the first.

When the sum of the *G* column has been obtained the rank correlation may be found from the formula

$$R = 1 - \frac{6\Sigma G}{(N^2 - 1)} \quad (96)$$

(Spearman's "foot-rule" correlation coefficient *R*,
found by the method of gains)

Substituting for ΣG its value of 10.5, and for *N* its value of 12, we obtain an *R* of .56. By means of Table 47 this *R* may be converted into an equivalent product-moment *r* of .79. It will be noted that this transmuted value of *r* compares favorably with the *r* (found from *p*) of .81.

The foot-rule formula gives only a rough estimate of the correlation between two sets of ranks; and is, in general, much less accurate than the rank-difference formula. The coefficient *R* "has a large, though except in the case of zero correlation, not definitely known *PE*; does not vary between - 1.00 and + 1.00; is not at all comparable in meaning with a product-moment coefficient; and in general has none of the merits

* Since the rankings run in order from 1 to 12, a rank of 6 is taken as higher (nearer 1) than a rank of 7.5.

TABLE 47

A TABLE TO INFER THE VALUE OF r FROM A GIVEN VALUE OF R

R	r	R	r	R	r	R	r
.00	.000						
.01	.018	.26	.429	.51	.742	.76	.937
.02	.036	.27	.444	.52	.753	.77	.942
.03	.054	.28	.458	.53	.763	.78	.947
.04	.071	.29	.472	.54	.772	.79	.952
.05	.089	.30	.486	.55	.782	.80	.956
.06	.107	.31	.500	.56	.791	.81	.961
.07	.124	.32	.514	.57	.801	.82	.965
.08	.141	.33	.528	.58	.810	.83	.968
.09	.158	.34	.541	.59	.818	.84	.972
.10	.176	.35	.554	.60	.827	.85	.975
.11	.192	.36	.567	.61	.836	.86	.979
.12	.209	.37	.580	.62	.844	.87	.981
.13	.226	.38	.593	.63	.852	.88	.984
.14	.242	.39	.606	.64	.860	.89	.987
.15	.259	.40	.618	.65	.867	.90	.989
.16	.275	.41	.630	.66	.875	.91	.991
.17	.291	.42	.642	.67	.882	.92	.993
.18	.307	.43	.654	.68	.889	.93	.995
.19	.323	.44	.666	.69	.896	.94	.996
.20	.338	.45	.677	.70	.902	.95	.997
.21	.354	.46	.689	.71	.908	.96	.998
.22	.369	.47	.700	.72	.915	.97	.999
.23	.384	.48	.711	.73	.921	.98	.9996
.24	.399	.49	.721	.74	.926	.99	.9999
.25	.414	.50	.732	.75	.932	1.00	1.0000

except brevity, of the formula based on the squares of differences in rank."* The foot-rule formula can be used to best advantage, perhaps, when the data are so meager or so crude as to make more precise calculation a waste of time. The formula may also be used in a preliminary survey to determine whether there is sufficient evidence of correlation to warrant the application of another method.

3. Summary of the Rank Methods

The product-moment method takes account of the size of a score as well as of its position in the series. Rank methods, on the other hand, take account *only* of the positions of the items or scores in the series. No account is taken of the size of the gap between adjacent scores. Individuals, for example,

* Kelley, T. L., *Statistical Method*, 1923, p. 193.

who score 90, 89, and 70 on a given test are ranked 1, 2, and 3 in order of merit, although the difference between 90 and 89 is 1, and the difference between 89 and 70 is 19. Considerable accuracy may also be lost in translating scores over into ranks. This is because gaps will appear in the rankings when a number of scores, all of the same size, receive the same rank. Rank methods are rarely used with test scores when N is larger than 30. Of the two rank methods, that of rank-differences gives more accurate results, and is much to be preferred to the "foot-rule."

II METHODS OF MEASURING CORRELATION OR ASSOCIATION WHEN THE DATA ARE GROUPED INTO CLASSES OR CATEGORIES

1. Bi-serial Correlation

In many problems it becomes important to calculate the correlation between traits or attributes, when the members of the group can be measured (i.e., given scores) in the first variable, but can only be classified into *two* categories in the second or "dichotomous" variable. (The term dichotomous means "cut into two parts.") We may, for instance, wish to know the correlation between MA and "social adjustment" in a group of nursery school children, when our subjects have been given scores in the first trait, but are simply classified as "socially adjusted" or "not socially adjusted" in the second trait. Other examples of dichotomous classification with reference to some attribute are athletic-non-athletic, Negro-white, radical-conservative, socially minded-mechanically minded, above eighth grade in school-below eighth grade, and the like. Many test and questionnaire items also are scored so as to give responses which fall into two categories; as, for example, problems marked Passed or Failed, statements marked True or False, personality inventory items answered Yes or No, interest test items marked Liked or Disliked, and so on. The correlation between a set of scores and a two-category classification (like those listed above) cannot be found by the ordinary product-moment formula or

by the rank methods. However, if we can assume that the attribute, for which we have made a two-way or dichotomous classification, would be continuous and normally distributed if more information were available so that classification could be made in finer units or steps, the correlation between such a trait and a set of scores may be computed by the *bi-seral correlation method*.

The calculation of bi-seral r is illustrated in Table 48. The problem is to find the correlation between total scores on a test and the answers to a single item in the test (Item 72); or put differently, to find whether those who make high scores on the test tend to answer Item 72 "Yes" more often than "No." The first column of Table 48 gives the steps of the score distribution. Column 2 gives the distribution of scores made by

TABLE 48

TO ILLUSTRATE THE CALCULATION OF THE BI-SERIAL r
Between Total Scores on a Test and the Answers to a
Single Item on the Test

Scores on Test	Responses to Item 72 "Yes" "No"		f	
				$M = 58.05$; mean of all scores ($N = 100$)
				$\sigma = 11.63$; σ of all scores ($N = 100$)
80-84	3		3	$M_p = 60.08$; mean of "Yes" responses
75-79	4	2	6	($N = 60$)
70-74	6	2	8	$M_q = 55.00$; mean of "No" responses
65-69	5	5	10	($N = 40$)
60-64	10	9	19	
55-59	10	5	15	$p = 60$; percent answering "Yes" to
50-54	15	5	20	Item 72
45-49	4	3	7	$q = .40$; percent answering "No" to
40-44	3	2	5	Item 72
35-39		4	4	$z = .386$; height of ordinate separating
30-34		2	2	60% from 40% in a normal
25-29		1	1	distribution (Table 49)
	60	40	100	
	(p)	(q)		

$$r_{bis} = \frac{M_p - M_q}{\sigma} \cdot \frac{pq}{z} \quad (97) \quad PE_{r_{bis}} = \frac{.6745 \left(\frac{\sqrt{pq}}{z} - r_{bis}^2 \right)}{\sqrt{N}} \quad (98)$$

$$= \frac{60.08 - 55.00}{11.63} \times \frac{(.60)(.40)}{.386} \quad = \frac{.6745 \left(\frac{\sqrt{.24}}{.386} - (.27)^2 \right)}{\sqrt{100}}$$

$$= .27 \pm .08 \quad = .08$$

the 60 subjects who answered "Yes" to Item 72, and column 3 the distribution of scores made by the 40 subjects who answered "No." The sum of all of the frequencies on the score-intervals gives the total distribution of 100 cases (see column 4). The steps in calculating bi-serial r from here on are as follows:

Step 1. Calculate M_p , the mean of the scores made by the 60 subjects who answered "Yes" to Item 72. Also calculate M_q , the mean of the scores made by the 40 subjects who answered "No" to Item 72. In our problem, $M_p = 60.08$, and $M_q = 55.00$.

Step 2. Calculate the σ of the whole distribution — the distribution of the 100 scores.* This σ , which equals 11.63, gives the spread of the test scores in the entire group.

Step 3. Sixty percent of the group (p) answered "Yes" to Item 72, and forty percent (q) answered "No" (p always equals $1 - q$). Assuming a normal distribution of opinion on this item (varying from complete agreement on through indifference to complete disagreement) upon which a dichotomous division has been forced, we place the dividing line between the "Yes" and "No" groups at a distance of ten percent from the middle of the curve, as shown in the figure below.

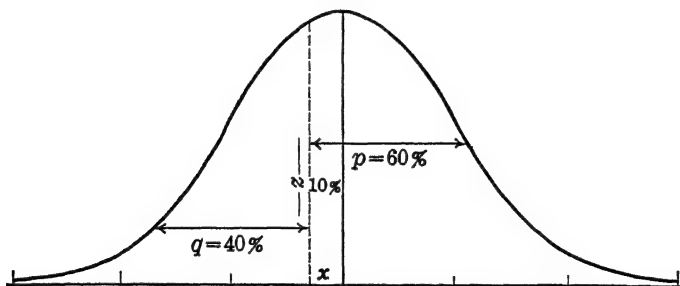


FIG. 55.

From Table 49, the height of the ordinate (i.e., z) which is ten percent removed from the mean of a normal distribution is .386.

* If the grouping is in broad intervals (e.g., less than 12) Sheppard's correction (p. 43) should be applied to the calculated σ .

Step 4. Having computed M_p , M_q , σ , p , q , and z , we find r_{bis} from the formula

$$r_{bis} = \frac{M_p - M_q}{\sigma} \cdot \frac{pq}{z} \quad (97)$$

(bi-serial coefficient of correlation or bi-serial r)

in which, as illustrated by the problem above, and shown in Table 48

M_p = mean of the group in the first category (usually the group showing superior or more desirable characteristics)

M_q = mean of the group in the second category

σ = standard deviation of the entire group

p = percent of the whole group in category one

q = percent of the whole group in category two ($p = 1 - q$)

z = height of the ordinate in the normal curve dividing p from q .

In Table 48, r_{bis} is .27, indicating a tendency, though not a strong one, for "Yes" answers to Item 72 to accompany high total scores.

The PE of r_{bis} is given by the formula

$$PE_{r_{bis}} = \frac{.6745 \left(\frac{\sqrt{pq}}{z} - r_{bis}^2 \right)}{\sqrt{N}} \quad (98)$$

(PE of r_{bis} for values of p and q greater than .05)

From this formula the PE of r_{bis} of .27 in Table 48 is found to be .08, which is slightly higher than the PE of a product-moment r of .27.

There is another — and slightly different — formula for bi-serial r which is often useful. This is

$$r_{bis} = \frac{M_p - M_T}{\sigma} \cdot \frac{p}{z} \quad (99)$$

(bi-serial coefficient of correlation or bi-serial r in terms of M_T , the mean of the total group)

in which

M_p = mean of the group in the first (or p) category

M_T = mean of entire group

σ = standard deviation of entire group

p = percent of whole group in category one

z = height of ordinate in normal curve dividing p from q .

Substituting in formula (99) the values for M_p , M_T , σ , p , and z , shown in Table 48, we have

$$r_{bis} = \frac{60.08 - 58.05}{11.63} \times \frac{.60}{.386} = .27$$

which checks our previous result.

TABLE 49

DEVIATES (x/σ) IN TERMS OF σ -UNITS AND ORDINATES (z) FOR
GIVEN AREAS MEASURED FROM THE MEAN OF A NORMAL
DISTRIBUTION WHOSE TOTAL AREA = 1.00

$[x/\sigma = x]$					
Area from the Mean (α)	x or (x/σ)	z	Area from the Mean (α)	x or (x/σ)	z
.00	.000	.399	.26	.706	.311
.01	.025	.399	.27	.739	.304
.02	.050	.398	.28	.772	.296
.03	.075	.398	.29	.806	.288
.04	.100	.397	.30	.842	.280
.05	.126	.396	.31	.878	.271
.06	.151	.394	.32	.915	.262
.07	.176	.393	.33	.954	.253
.08	.202	.391	.34	.995	.243
.09	.228	.389	.35	1.036	.233
.10	.253	.386	.36	1.080	.223
.11	.279	.384	.37	1.126	.212
.12	.305	.381	.38	1.175	.200
.13	.332	.378	.39	1.227	.188
.14	.358	.374	.40	1.282	.176
.15	.385	.370	.41	1.341	.162
.16	.412	.366	.42	1.405	.149
.17	.440	.362	.43	1.476	.134
.18	.468	.358	.44	1.555	.119
.19	.496	.353	.45	1.645	.103
.20	.524	.348	.46	1.751	.086
.21	.553	.342	.47	1.881	.068
.22	.583	.337	.48	2.054	.048
.23	.613	.331	.49	2.326	.027
.24	.643	.324	.50	∞	.000
.25	.675	.318			

Formula (99) is especially well suited to those cases in which sub-groups having different characteristics are drawn out of the total group for study, the total group remaining the same.

Such a situation is illustrated by the case in which success and failure upon individual items in a questionnaire are correlated against total score on the same test. The advantage of formula (99) is that the M and σ of the total group having once been computed, only the p , M_p and z of each additional sub-group are needed in order to calculate r_{bis} .

2. Tetrachoric Correlation

As we have seen in the last section, when one variable is continuous and is expressed in the form of test scores, and the other is dichotomous and is expressed by a two-fold classification, bi-serial r may be employed to estimate the correlation between the two variables. A problem which is still more specialized than that to which bi-serial r is applicable presents itself when *both* variables are dichotomous. A situation of this sort yields a 2×2 fold correlation table, from which a modified form of the product-moment coefficient, called *tetrachoric r* , may be calculated. Tetrachoric r is useful in psychology and education when one wishes to calculate the correlation between two characters or attributes neither of which is directly measurable, but both of which are capable of being grouped into at least two categories. Thus, if we wish to measure the correlation between school attendance and employment, persons might be classified, on the one hand, into those who have attended high school and those who have not; and, on the other hand, into those who are employed and those who are unemployed. Or if we wish to know the correlation between intelligence and social maturity, children might be classified as "above average" and "below average" in intelligence, on the one hand, and as socially mature and socially immature on the other. The tetrachoric correlation method assumes that the two variables being studied are essentially continuous, and would be found to be normally distributed, if it were possible to classify them more exactly into finer groupings.

Table 50 illustrates a 2×2 fold table, and shows the steps involved in calculating tetrachoric r . The problem is to find

TABLE 50

TO ILLUSTRATE THE CALCULATION OF TETRACHORIC r (r_z)
(The data are hypothetical)

Y-variable	X-variable		
	100 Salesmen		Totals
	Unsuccessful	Successful	
Socially Well Adjusted	25 (a)	35 (b)	60 $p = 60\%$
Socially Poorly Adjusted	30 (c)	10 (d)	40 $q = 40\%$
Totals	55 $q' = 55\%$	45 $p' = 45\%$	100

For $p = .60, q = .40, \alpha = .10$
 $x = -.253$
 $z = .386$ (Table 49)

For $p' = .45, q' = .55, \alpha = .05$
 $x' = .126$
 $z' = .396$ (Table 49)

$$\frac{bc - ad}{N^2 z z'} = r + \frac{xx' r^2}{2} \quad (101)$$

$$\frac{1050 - 250}{100^2 (.386)(.396)} = r + \frac{(-.253)(.126)r^2}{2}$$

$$.523 = r - .016r^2$$

or

$$.016r^2 - r + .523 = 0^*$$

$$r = \frac{+1 \pm \sqrt{1 - 4(.016)(.523)}}{2 \times .016} = \frac{+1 \pm \sqrt{1 - .033472}}{.032}$$

$$= \frac{+1 \pm .9831}{.032}$$

$$= .53 \text{ (taking numerator as } +1 - .9831)$$

$$= +62 \text{ (taking numerator as } +1 + .9831)$$

* The general form of a quadratic equation is $ax^2 + bx + c = 0$. The two values of x (i.e., the roots of the equation) may be computed by the formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

In the equation $.016r^2 - r + .523 = 0$, $a = .016$; $b = -1.00$; and $c = .523$. Hence,

$$r = \frac{+1 \pm \sqrt{1 - 4(.016)(.523)}}{2 \times .016}$$

$$= .53 \text{ or } 62 \text{ (an impossible value)}$$

whether salesmen who are successful tend to be "socially well adjusted" more often than salesmen who are unsuccessful. The data are hypothetical. The X -variable (along the top of the diagram) is divided into two categories "successful" and "unsuccessful"; and the Y -variable (along the left of the diagram) is divided into two categories "socially well adjusted" and "socially poorly adjusted." The sums of the rows show that 60 salesmen ($a + b$) out of the sample of 100 are classed as well adjusted socially, and that 40 salesmen ($c + d$) are classed as poorly adjusted socially.* The proportions in each category (p and q) are 60% and 40%, respectively. The sums of the columns show that 55 of the 100 salesmen are classified as unsuccessful, and 45 as successful; the proportions are 55% (q') and 45% (p'). On the assumption that "success in salesmanship" is distributed normally, from the proportions $p = .60$, and $q = .40$, we obtain an $x = -.253$, and $z = .386$. These last two values are read from Table 49 as follows. The perpendicular line (i.e., the ordinate, z) separating the *upper* 60% from the *lower* 40% in a normal curve is just 10% from the mean. Hence, entering the first column of Table 49 with $\alpha = .10$, we read $x = -.253$, and $z = .386$. See diagram below.

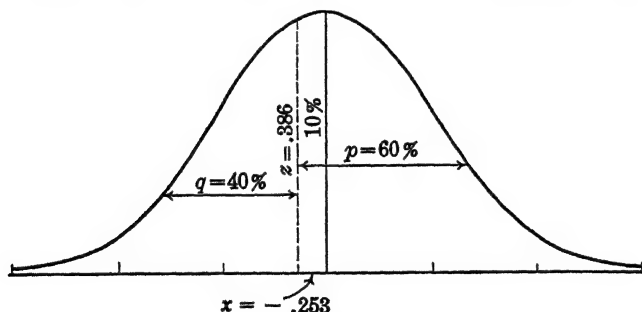


FIG. 56

* To accord with the plan of the ordinary correlation table (p. 279), the categories in Table 50 have been so arranged that concentration of data in the first and third quadrants denotes positive correlation; concentration of data in the second and fourth quadrants negative correlation.

The x' and z' values corresponding to $p' = .45$ and $q' = .55$ are calculated in the same way. The perpendicular line dividing the upper 45% (the percent successful) from the lower 55% (the percent unsuccessful) is 5% from the mean; and from Table 49, for $\alpha = .05$, $x' = .126$, and $z' = .396$. See diagram below.

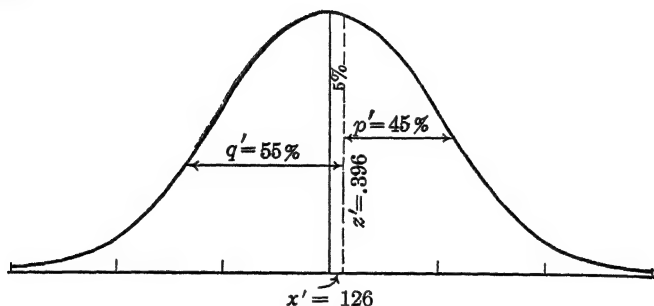


FIG. 57.

The formula for tetrachoric r may now be written as follows:

$$\frac{bc - ad}{N^2 z z'} = r_t + \frac{x x' r_t^2}{2!} + \frac{(x^2 - 1)(x'^2 - 1) r_t^3}{3!} + \frac{(x^3 - 3x)(x'^3 - 3x') r_t^4}{4!} \quad (100)$$

(formula for tetrachoric r to four terms) *

in which the x 's and x ''s represent σ -distances from the mean to the points separating the percentage in the upper category from the percentage in the lower category; the z 's and z ''s represent the heights of the ordinates at the points of division; a , b , c , and d are the entries in the four cells, N is the size of the sample, and r_t is the tetrachoric coefficient of correlation.

Formula (100) is clearly too long and too complicated to be of much real service in correlational work. It may be simplified by dropping all terms on the right hand side of the equation

* Pearson, Karl, *On the Correlation of Characters not Quantitatively Measurable*, Philosophical Transactions, Royal Society of London, Series A, 1900, 195, pp. 1-47.

containing r_t 's beyond the second power. The contribution of this part of the equation is usually small, and is negligible when r_t is low. Dropping terms beyond r^2 reduces formula (100) to the following quadratic equation —

$$\frac{bc - ad}{N^2 z z'} = r_t + \frac{xx' r_t^2}{2} \quad (101)$$

(approximate formula for tetrachoric r)

which may be readily solved for r_t .

In Table 50, bc is found to equal 1050, and ad to equal 250. Substituting these quantities, and for x , x' , z , z' and N^2 in formula (101), we obtain, as shown by the calculations in the table, $r_t = .53$. This coefficient indicates a substantial correlation between success in salesmanship and good social adjustment. In order to compute our r_t it is necessary that we solve a quadratic equation. The method of carrying through this solution is given in Table 50 and in the footnote at the bottom of the table. Note that only the first of the two solutions for r_t is a possible value, as the second is greater than unity.

The investigator who finds it necessary to calculate many tetrachoric r 's may greatly shorten his work by using the charts devised by Thurstone and his co-workers.* These charts enable one to obtain a solution for r_t by graphic methods, without the necessity of solving formula (101).

The formula for PE_{r_t} is an exceedingly complex expression and is not reproduced here. The derivation of this formula will be found in books dealing to a greater extent than does the present volume with the mathematics of statistical theory.† The computation of PE_{r_t} may be greatly shortened by the use of Pearson's Tables XXIII and XXIV.‡ If a high degree

* Chesire, L., Saffir, M., Thurstone, L. L., *Computing Diagrams for the Tetrachoric Correlation Coefficient*, University of Chicago Bookstore, 1933.

† See: Kelley, T. L., *Statistical Method*, 1923, pp. 253-258

Peters, C. C., and VanVoorhis, W. R., *Statistical Procedures and Their Mathematical Bases*, 1935, pp. 261-269

‡ Pearson, Karl, *Tables for Statisticians and Biometricians*, 1914, Introduction, xl-xli, and p. 35.

TABLE 51
TO ILLUSTRATE THE USE OF TETRACHORIC r IN EVALUATING
A GIVEN TEST
 $N = 125$
X-variable

Y-variable	College Juniors		
	Non-Science Majors	Science Majors	
Above Test Mean	24% (a)	35% (b)	$p = 59\%$
Below Test Mean	29% (c)	12% (d)	$q = 41\%$
	$q' = 53\%$	$p' = 47\%$	100%

For $p = .59, q = .41$ $x = -.228$ $z = .389$ For $p' = .47, q' = .53$ $x' = .075$ $z' = .398$

$$\frac{.1015 - .0288}{(.389)(.398)} = r + \frac{(-.228)(.075)r^2}{2} \quad (101)$$

$$.470 = r - .009r^2$$

$$\text{or} \quad .009r^2 - r + .470 = 0$$

$$r = \frac{+1 \pm \sqrt{1 - 4(.009)(.470)}}{2(.009)}$$

$$= \frac{+1 \pm .9915}{.018}$$

$$= .47, \text{ or } 111 \text{ (an impossible value)}$$

of accuracy is not demanded by the problem, an approximation to the PE of a tetrachoric r may be found in the following way. The PE_{r_t} is about 50% higher than the PE of an equivalent product-moment r , that is, the product-moment r equal to the given r_t and calculated from a sample of the same size as that upon which r_t is based. Hence, since the PE of a product-moment r of .53 is .05, for $N = 100$ (see Table 41), the PE of a tetrachoric r of .53 is approximately $.05 \times 1.5$ or .08.

Tetrachoric r may often be usefully employed as a means of evaluating a given test or questionnaire in accordance with its ability to separate two contrasted or "criterion" groups. An example is given in Table 51 (the data are hypothetical). The

problem is to find whether a test of deductive reasoning (e.g., a syllogism test) will differentiate 59 college juniors majoring in science from 66 college juniors majoring in literature or languages (non-science). The X -variable is divided into science majors and non-science majors; the Y -variable into those above and those below the mean of the test, i.e., the mean score established by the entire junior class. The entries in the cells, a , b , c , and d , are expressed in percents, so that N^2 in formula (101) is 1.00. As shown in Table 51, the correlation between majoring in science and high scores on the syllogism test is $.47 \pm .07$ (approximate PE). If one were investigating a number of tests with a view toward determining their relative value as indicators of scientific aptitude, the worth of each test could be evaluated in accordance with its ability to separate the two criterion groups *

3. The Chi-Square (χ^2) Test

We have already had occasion (in Chapter V) to use the Chi-Square Test as a means of discovering whether a given frequency distribution departs significantly from the normal form. In addition to providing a precise measure of "goodness of fit," χ^2 is applicable to a variety of problems in psychology and education. For example, χ^2 is often very useful for testing whether certain experimentally obtained results differ significantly from those to be expected by "chance"; or whether obtained results agree or disagree with the findings to be expected on some other hypothesis. The χ^2 method differs from the correlational methods hitherto described in that it does not yield a *coefficient* which gives a measure of *degree* of association or relationship. What χ^2 does is to provide a measure of the *probability* that two sets of data are dependent (definitely associated) or are independent (significantly different). Several illustrations of this use of χ^2 will be given in this section.

* For a discussion of the application of tetrachoric r to problems involving two widely separated or extreme groups in which the middle group is eliminated, see Peters, C C, and VanVoorhis, W. R., *Statistical Procedures and Their Mathematical Bases*, 1935, pp. 269-278.

- (1) Comparison of observed results and those expected by "chance"

The meaning of χ^2 as used in this section* may be made clearer, perhaps, if we rewrite formula (21) given on page 123. Let o = the frequency of occurrence of some fact actually *observed* or determined experimentally; and let e = the *expected* frequency of occurrence of the same fact on the hypothesis of chance (or some other hypothesis). Then if $x = (o - e)$

$$\chi^2 = \Sigma \left(\frac{x^2}{e} \right) \quad (102)$$

(χ^2 formula for testing agreement between observed and expected results)

that is, the sum of all of the differences between observed and expected results, *squared*, and divided by the number expected in each case. The more closely the observed results agree with the expected, the smaller χ^2 , and the greater the probability of agreement between the two sets of data compared. On the other hand, the larger χ^2 , the greater the probability of a true divergence of experimentally observed from theoretical results. To evaluate χ^2 , we enter Table 52 with the given value of χ^2 , and n . The quantity n equals $(r - 1)(c - 1)$, r being the number of rows, and c the number of columns in which the data are tabulated. From the given values of χ^2 and n is found P , the probability that the given divergence between observed and theoretical results is significant and cannot be explained "by chance."

Table 53 illustrates the application of the χ^2 test to a fairly common problem in questionnaire construction. Forty-eight subjects were asked to express their attitude toward the proposition "Should the United States join the League of Nations?" by marking F (favorable), I (indifferent) or U (unfavorable). The experimentally obtained results are given in the first row opposite "Obtained (o)" — 24 are favorable, 12 are indifferent,

* For a comprehensive discussion of χ^2 as applied to various problems, see Fisher, R. A., *Statistical Methods for Research Workers*, 1930, Chapters 3 and 4.

TABLE 52. Giving the Probability (P) that, with a Given n , the χ^2 Value Obtained in the Comparison of the Distribution of a Sample with that of a Theoretical Series Indicates that the Sample Belongs to or Has Arisen Out of Such Series. (The values of χ^2 are printed in the body of the table.) For larger values of n , the expression $\sqrt{2n} - \sqrt{2n-1}$ may be used as a normal deviate with unit standard error.

Adapted from R. A. Fisher's *Statistical Method for Research Workers*, Oliver & Boyd, by permission of publishers

n	$P = 0.99$	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0.000157	0.000628	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.685	4.642	6.251	7.815	9.637	11.341
4	0.297	0.429	0.711	1.064	1.649	2.195	3.557	4.878	5.989	7.779	9.488	11.668	13.277
5	0.554	0.762	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	0.872	1.134	1.635	2.204	3.072	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.668
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.007	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.380	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.348	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.441	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.668	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.439	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.870	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.555	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.236	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.692	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.266	43.773	47.968	50.892

TABLE 53

TO ILLUSTRATE THE USE OF THE χ^2 FORMULA IN TESTING THE AGREEMENT BETWEEN OBTAINED AND EXPECTED RESULTS

Data represent actual responses of 48 subjects to a proposition and the responses expected by chance

	Answers			Total
	Favorable	Indifferent	Unfavorable	
Obtained (<i>o</i>)	24	12	12	48
Expected (<i>e</i>)	16	16	16	48
$x(o - e)$	8	4	4	
x^2	64	16	16	
$\frac{x^2}{e}$	4	1	1	6
$\chi^2 = \Sigma\left(\frac{x^2}{e}\right) = 6.0$	$n = 2$		$P = .05$ (Table 52)	

and 12 are unfavorable. In the second row opposite "Expected (*e*)" is the distribution of answers to be expected by chance; the results to be expected if each alternative is selected equally often, i.e., without preference. In the third row are tabulated the x 's (the excess of *o* over *e*). When each x^2 has been divided by its *e*, i.e., by the *e* of its column, $\Sigma\left(\frac{x^2}{e}\right)$ gives a

χ^2 which equals 6.0. There are three columns and two rows in the table; hence $n = (3 - 1)(2 - 1)$ or 2. Entering Table 52 with a $\chi^2 = 6.0$, and $n = 2$, we obtain a P equal to .05*. This means that there are only 5 chances in 100 that the obtained χ^2 which represents the divergence of our experimentally obtained entries from the expected entries, could have arisen "by chance."† There are 95 chances in 100, therefore, that the distribution of obtained answers differs *significantly* from the distribution of answers to be expected on the hypothesis of equal distribution of answers (no preference). We may be

* A P of .02 or less may be taken as indicative of a significant deviation from expectancy

† To be strictly accurate this sentence should read . . . "there are only 5 chances in 100 that the obtained χ^2 or a greater value could have arisen from fluctuations of random sampling"

reasonably sure, then, that this group favors the given proposition.

TABLE 54

THE χ^2 TEST APPLIED TO THE DISTRIBUTIONS OF OBSERVED
AND CHANCE RESPONSES TO A 5-RESPONSE ITEM

	Answers					Total
	Strongly Approve	Approve	Indiffer- ent	Disap- prove	Strongly Disap- prove	
Obtained (<i>o</i>)	23	18	24	17	18	100
Expected (<i>e</i>)	20	20	20	20	20	100
$x(o - e)$	3	2	4	3	2	
x^2	9	4	16	9	4	
$\frac{x^2}{e}$.45	.20	.80	.45	.20	2.10
$\chi^2 = 2.10$	$n = 4$		$P = .70$ (Table 52)			

A second example of the Chi-Square Test is given in Table 54. The data here represent the responses to an item on an attitude scale. Five answers were allowed: Strongly Approve, Approve, Indifferent, Disapprove, Strongly Disapprove. The distribution of obtained answers for 100 individuals is given in the first row, and the responses to be expected (if there is no preference) in the second row. $\chi^2 = 2.10$, and $n = (5 - 1)(2 - 1)$ or 4. From Table 52, a $\chi^2 = 2.10$, and an $n = 4$, give a $P = .70$ (approximately). Hence there are 70 chances in 100 that the given value of χ^2 (that is, the discrepancy between the experimentally obtained and the theoretically expected responses) could have arisen by chance; and only about 30 chances in 100 that "real" factors were at work to produce the difference. There is no evidence, then, of a marked preference either for agreement or disagreement with the given item; the divergence of the obtained answers from a chance distribution is not significant.

(2) Contingency tables and tests of independence

We have seen above how χ^2 may be employed to test the agreement of actually obtained and theoretically expected re-

sults. Other useful applications of χ^2 , now to be described, are (1) in testing the hypothesis of "independence" in a contingency table; and (2) in testing whether the distributions of answers, scores, or expressed attitudes obtained from two different groups are — or are not — significantly divergent.

Table 55 is a "contingency table," i.e., a table in which is represented the joint possession by two groups of varying degrees (often expressed on a qualitative scale) of some attribute, character, or other trait. In the tabulation in Table 55, 730 fathers and 730 sons are grouped conjointly, with respect to temperament, under four heads. Reading down the first column, we find that out of 260 fathers characterized as "merry" in temperament, 122 have "merry" sons, 10 have "melancholy" sons, 70 have sons whose temperaments are "alternating," and 58 have "even-tempered" sons. Taking the first row, we find 278 "merry" sons of whom 122 have "merry" fathers, 8 have "melancholy" fathers, 81 have fathers whose moods are "alternating," and 67 have "even-tempered fathers."

Before we can compute χ^2 , we must calculate the "independence value" for each cell in the contingency table. These values, which are represented by the figures in parentheses in the different cells, give the number of father and son pairs which we should expect to find in the various cells in the absence of any *actual association* between temperamental characteristics in father and son. The method of calculating "independence values" is shown in Table 55. To illustrate, there are 260 fathers and 278 sons who are described as merry. If there were *no association* at all between merriness in father and son,

we should expect to find at least $\frac{260 \times 278}{730}$ or 99 merry fathers

with merry sons by the operation of chance alone. This "chance value" is obtained as follows. We have found that 278/730 of *all* of the sons are described as merry. This percentage should hold for the sons of *all* fathers if there is no dependence of son on father with respect to temperament. Hence, 278/730 of the 260 merry fathers, i.e., 99, should have merry sons even if

TABLE 55

COMPARISON OF FATHERS AND SONS WITH RESPECT
TO TEMPERAMENT *

(To Illustrate the Computation of χ^2 from a Contingency Table)

Fathers

	Merry	Melan- choly	Alternat- ing	Even	Totals
Merry	(99 0) 122	(9.5) 8	(97 1) 81	(72 3) 67	278
Melancholy	(10 3) 10	(1.0) 2	(10.1) 7	(7 5) 10	29
Alternating	(88 3) 70	(8 5) 9	(86 6) 101	(64 5) 68	248
Even	(62 3) 58	(6 0) 6	(61.1) 66	(45 5) 45	175
Totals	260	25	255	190	730

I. Independence Values

$$\begin{array}{llll}
 \frac{260 \times 278}{730} = 99.0 & \frac{25 \times 278}{730} = 9.5 & \frac{255 \times 278}{730} = 97.1 & \frac{190 \times 278}{730} = 72.3 \\
 \frac{260 \times 29}{730} = 10.3 & \frac{25 \times 29}{730} = 1.0 & \frac{255 \times 29}{730} = 10.1 & \frac{190 \times 29}{730} = 7.5 \\
 \frac{260 \times 248}{730} = 88.3 & \frac{25 \times 248}{730} = 8.5 & \frac{255 \times 248}{730} = 86.6 & \frac{190 \times 248}{730} = 64.5 \\
 \frac{260 \times 175}{730} = 62.3 & \frac{25 \times 175}{730} = 6.0 & \frac{255 \times 175}{730} = 61.1 & \frac{190 \times 175}{730} = 45.5
 \end{array}$$

II. Calculation of χ^2 .

$$\begin{array}{lll}
 (23)^2 \div 99 = 5.34 & (5)^2 \div 8.5 = .03 & (4.9)^2 \div 61.1 = .39 \\
 (-3)^2 \div 10.3 = .01 & (0)^2 \div 6.0 = .00 & (-5.3)^2 \div 72.3 = .39 \\
 (-18.3)^2 \div 88.3 = 3.79 & (-16.1)^2 \div 97.1 = 2.67 & (2.5)^2 \div 7.5 = .83 \\
 (-4.3)^2 \div 62.3 = .30 & (-3.1)^2 \div 10.1 = .95 & (3.5)^2 \div 64.5 = .20 \\
 (-1.5)^2 \div 9.5 = .24 & (14.4)^2 \div 86.6 = 2.39 & (-.5)^2 \div 45.5 = .01 \\
 (1.0)^2 \div 1.0 = 1.00 & & \\
 \chi^2 = 18.54 & n = 9 & P = .03
 \end{array}$$

* Quoted from Brown, W., and Thomson, G., *Essentials of Mental Measurement*, 1925, p. 125.

there were *no* connection between temperament in father and son. Independence values for all of the cells have been found in the same way, as shown in Table 55.

When expected or independence values have been computed for each cell, we find the difference between the observed and expected values in *each* cell, square these differences, and divide in each instance by the given expected or independence value. The sum of these quotients by formula (102) gives χ^2 . In the present problem, $\chi^2 = 18.54$, and $n = (4 - 1)(4 - 1)$ or 9 (p. 378). Hence, P , by interpolation in Table 52, is about .03. This result means that 3 times in 100 we should expect to get by chance a χ^2 equal to or greater than 18.54. The given P does not quite pass our criterion of significance (namely, .02). But the indications are certainly strong that the association between father and son with respect to temperament is too great to be accounted for *entirely* by fortuitous influences (chance association).

A more detailed analysis of the association exhibited in Table 55 is provided by Table 56, in which are entered the contributions made to the total χ^2 by each cell separately. Columns 1 and 3 show the strongest tendency toward a greater-than-chance association, i.e., have largest χ^2 values. Going to the separate cells, we find a decided tendency for sons of merry fathers to be classified also as merry, and a somewhat lesser tendency for the sons of merry fathers to be classified as "alternating" in temperament. The sons of melancholy fathers are classified as merry, melancholy, alternating, and even about as often as chance would allow. In general, it appears that the sons of fathers described as merry and alternating in temperament tend to be put into the same descriptive categories as their fathers more often than is true of the sons of melancholy and even-tempered fathers. Analysis of a contingency table in terms of the specific contribution to χ^2 made by each cell may often be enlightening. In the first place, we are enabled to see just where the discrepancy between obtained and expected results lies. Furthermore, we may gain insight into the ques-

TABLE 56
SHOWING χ^2 VALUES FOR EACH CELL IN TABLE 55
Fathers

Sons		Merry	Melan- choly	Alternat- ing	Even	Totals
	Merry	5 34	.24	2.67	.39	8.64
	Melancholy	01	1 00	95	.83	2.79
	Alternating	3 79	.03	2 39	.20	6.41
	Even	30	00	39	.01	.70
	Totals	9 44	1.27	6 40	1.43	18 54

tions of which categories are the more often used, and hence are, perhaps, the more objective; whether certain groups are producing most of the discrepancy between hypothetical and experimental results, and so on.

The application of χ^2 to another problem, that of ascertaining whether two groups differ significantly in their responses to an item of a questionnaire, say, is illustrated in Table 57. The data in the table represent the responses to the question "Are you easily discouraged?" given by 90 normal and 100 abnormal subjects. Three alternative answers are allowed: Yes, No, and ? (uncertain). It is clear from Table 57 that, while the normals and abnormals answer No and ? about equally often, nearly twice as many abnormals as normals answer Yes. Is this greater frequency of Yes responses by abnormals sufficient to make this question valuable as a means of differentiating between the two groups?

The answer to this question is given in Table 57. Since we are not primarily concerned in this problem with the theoretical or expected results, a somewhat different and shorter method of calculating χ^2 may be employed. This method is as follows. Let a and a' be any pair of obtained column entries and N and N' be the corresponding row totals. Then the contribution to χ^2 of any given column is

$$\frac{1}{a + a'} (aN' - a'N)^2$$

TABLE 57

TO ILLUSTRATE THE χ^2 TEST WHEN THE PROBLEM IS TO ESTIMATE
THE SIGNIFICANCE OF THE DIFFERENCES BETWEEN TWO
SETS OF ANSWERS TO A QUESTIONNAIRE ITEM

	Answers to Given Item			Totals
	Yes	No	?	
Normals	14	66	10	90 (<i>N</i>)
Abnormals	27	66	7	100 (<i>N'</i>)
	41	132	17	190

$$\frac{1}{41}(14 \times 100 - 27 \times 90)^2 = 25875.6$$

$$\frac{1}{132}(66 \times 100 - 66 \times 90)^2 = 3300.0$$

$$\frac{1}{17}(10 \times 100 - 7 \times 90)^2 = \frac{8052.9}{37228.5}$$

$$\chi^2 = \frac{37228.5}{90 \times 100} = 4.14$$

$n = 2$ P lies between .20 and .10 (Table 52)

For column 1, wherein $a = 14$, $a' = 27$, $N = 90$, and $N' = 100$, the contribution to χ^2 is

$$\frac{1}{41}(14 \times 100 - 27 \times 90)^2 \text{ or } 25875.6$$

The contribution to the total χ^2 of each of the other columns is found in the same way. In the second column, for example, $a = 66$, $a' = 66$, $N = 90$, and $N' = 100$, and the contribution of this column is 3300 (see Table 57 for calculations). The sum of the contributions of the three columns divided by NN' (90×100) gives a $\chi^2 = 4.14$. The term $n = (3 - 1)(2 - 1)$ or 2. Entering Table 52 with $\chi^2 = 4.14$, and $n = 2$, we find that P lies between .20 and .10. There is no conclusive evidence, therefore, that abnormals report themselves to be more easily discouraged than do normals. But since there are only about 15 chances in 100 that the differences between the two sets of answers can be attributed to chance, we may feel sure of a fairly strong trend toward greater reported discouragement in abnormals.

Summary

The chief advantage of the χ^2 test is that it provides a precise measure of the probability that observed and expected occurrences differ significantly. χ^2 varies with the size of the sample; i.e., the probability of association will become greater (or less) as N increases. The χ^2 test, therefore, may indicate significant association in a small sample, and negligible or non-significant association in a large sample. This shows the desirability of large groups when probability relations are to be considered.

The χ^2 test is applicable to data which have been grouped into frequencies; it should not be used with indices, ratios, or percents, since in the latter the size of the sample does not directly affect χ^2 . Perhaps the chief drawback to the χ^2 test is the fact that it does not provide an index of *degree* of relationship as does the product-moment r , but simply indicates the *probability* of association. χ^2 's from different tables are not strictly comparable. Contingency tables may be compared, however, in terms of the *coefficient of contingency*, a measure derived from χ^2 .

4. The Contingency Coefficient

The coefficient of *mean square contingency*, or more simply the *contingency coefficient*, was developed by Karl Pearson in 1904. The contingency coefficient, or C , is based upon χ^2 ; but it differs from χ^2 in that it provides a measure of correlation which under certain conditions (p. 391) is comparable to the product-moment r . C bears the following relation to χ^2 :

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \quad (103)$$

(a formula for C , the contingency coefficient in terms of χ^2)

Applying formula (103), we find that the C of Table 55 is $\sqrt{\frac{(18.54)}{730 + (18.54)}}$ or .16. Taken at face value, this C indicates a very small degree of relationship between temperament in father and son. To find whether the obtained C is indicative of

significant relationship, we must calculate its PE . Unfortunately, the PE of a C is a complex expression * and is somewhat tedious to compute. For a $C = .00$, however, $PE_C = \frac{.6745}{\sqrt{N}}$; and this formula may be employed to give at least a rough test of the significance of an obtained C . If C in Table 55 were .00, its PE would be $\frac{.6745}{\sqrt{730}}$ or .025. Our C of .16, therefore, is .16/.025 or about $6PE_C$ removed from a C of .00. Hence, $C = .16$ may be considered to indicate a small but significant degree of correlation between temperament in father and son. This result checks our conclusion from the χ^2 test on page 383.

When one is not interested in χ^2 itself, it is possible to compute C directly rather than by way of χ^2 . There are two methods of calculating C which will be given in order.

(1) Method I of Calculating C

In Table 58 the computation of C from a contingency table is shown. The table represents the association of father and son with respect to eye color. The independence values for each cell have been computed as shown above in Table 55. To illustrate again the method of calculation, 335/1000 of all sons are described as blue-eyed. This proportion of 335 (i.e., $\frac{335 \times 358}{1000}$) gives 120 as the number of fathers who might be expected to have blue-eyed sons "by chance," as contrasted with the 194 fathers who actually have blue-eyed sons. When each independence value has been found, we square each obtained cell entry, and divide by its own independence value as shown in Table 58. The sum of these quotients gives S ; and given S and N , C is calculated by the formula

$$C = \sqrt{\frac{S - N}{S}} \quad (104)$$

(formula for C , coefficient of contingency, calculated directly)

In Table 58, C is .46, which indicates a fairly strong degree of correlation between eye color in father and son

* See Kelley, T. L., *Statistical Method*, 1923, p. 269.

TABLE 58

TO ILLUSTRATE THE CALCULATION OF C , THE COEFFICIENT OF
MEAN SQUARE CONTINGENCY

		Father's Eye Color					II. Calculation of C	
		Blue	Gray	Hazel	Brown	Totals		
Son's Eye Color	Blue	(120) 194	(88) 70	(60) 41	(66) 30	335	$\frac{(194)^2}{120} =$	313.6
	Gray	(102) 83	(75) 124	(51) 41	(56) 36	284	$\frac{(83)^2}{102} =$	67.5
	Hazel	(49) 25	(36) 34	(25) 55	(27) 23	137	$\frac{(25)^2}{49} =$	12.8
	Brown	(87) 56	(64) 36	(44) 43	(48) 109	244	$\frac{(56)^2}{87} =$	36.0
Totals		358	264	180	198	1000	$\frac{(70)^2}{88} =$	55.7
							$\frac{(124)^2}{75} =$	205.0
							$\frac{(34)^2}{36} =$	32.1
							$\frac{(36)^2}{64} =$	20.3
							$\frac{(41)^2}{60} =$	28.0
							$\frac{(41)^2}{51} =$	33.0
							$\frac{(55)^2}{25} =$	121.0
							$\frac{(43)^2}{44} =$	42.0
							$\frac{(30)^2}{66} =$	13.6
							$\frac{(36)^2}{56} =$	23.1
							$\frac{(23)^2}{27} =$	19.6
							$\frac{(109)^2}{48} =$	247.5
							$S =$	1270.8
							$N =$	1000
							$S - N =$	270.8

$$C = \sqrt{\frac{S - N}{S}} = \sqrt{\frac{270.8}{1270.8}} = .46$$

C may be either plus or minus, the sign to be affixed depending upon an inspection of the contingency table itself. In Table 58 it is evident that pigmentation of eyes in father and son is positively correlated * and hence C is positive.

A disadvantage of the contingency coefficient is the fact that C does not remain constant, for the same data, when the number of classes varies. The C calculated from a 3×3 fold table will not ordinarily equal the C calculated from the same data arranged in, say, a 5×5 fold table. Moreover, the maximum value which C can take will depend upon the fineness of the classification employed. Yule † has shown that

when the number of classes = 2,	C cannot exceed .707
when the number of classes = 3,	C cannot exceed .816
when the number of classes = 4,	C cannot exceed .866
when the number of classes = 5,	C cannot exceed .894
when the number of classes = 6,	C cannot exceed .913
when the number of classes = 7,	C cannot exceed .926
when the number of classes = 8,	C cannot exceed .935
when the number of classes = 9,	C cannot exceed .943
when the number of classes = 10,	C cannot exceed .949

In the light of this table, Yule suggests that we "restrict the use of the 'coefficient of contingency' to 5×5 fold or finer classifications" in order that the maximum value of C may be as near unity as possible. At the same time, we should avoid a too-fine classification or C will be affected by slight or "casual irregularities of no physical significance"; and, in addition, the arithmetic of calculation will be greatly (and needlessly) increased. Pearson ‡ has worked out a correction for "broad

* We note, for example, that 194 blue-eyed fathers have blue-eyed sons, while only 30 brown-eyed fathers have blue-eyed sons. Moreover, 109 brown-eyed fathers have brown-eyed sons while only 56 blue-eyed fathers have brown-eyed sons. Comparisons of this sort will show that association between pigmentation in the eyes of father and son is positive.

† Yule, G. U., *An Introduction to the Theory of Statistics*, 1929, p. 66.

‡ Pearson, Karl, *On the Measurement of the Influences of "Broad Categories" on Correlation*, *Biometrika*, 1913, 9, p. 130; also see the discussion in Peters, C. C., and VanVoorhis, W. R., *Statistical Procedures and Their Mathematical Bases*, 1935, pp. 289-295.

categories" which should be applied to C 's calculated from 4×4 fold or broader groupings if C is to be compared with r . For 5×5 fold or finer classifications, this correction is so small that for practical purposes it may be disregarded.

Since the classification in Table 58 is 4×4 fold, the value of C will be increased if corrected for broad categories. An approximate correction, which is easier to apply than Pearson's correction, can be made by dividing the obtained C by the maximum value which C can take in a 4×4 fold contingency table. In the present problem, dividing our C of .46 by .866 (the maximum C for a 4×4 fold table) we obtain a corrected C of .53. This value may be taken as approximately equal to r ; it indicates a fairly high correlation between pigmentation of eyes in father and son.

As already indicated (p. 387) the relation of C to r is, under certain conditions, very close. C is practically equivalent to r when (1) the grouping is relatively fine — 5×5 fold or finer; (2) when the sample is large; and (3) when we know, or are justified in assuming, that the characters or attributes under investigation are normally distributed.

(2) Method II for Calculating C

The arithmetic involved in computing C may be lessened somewhat by combining the twofold process of (1) calculating independence values and (2) dividing the square of each cell frequency by its independence value. This method is illustrated in Table 59. The first occupied cell in the first column of the table has a frequency of 1 and an independence value of $\frac{99 \times 8}{384}$; hence the cell frequency squared and divided by the in-

dependence value is $\frac{1 \times 384}{8 \times 99}$. This fraction, namely, $\frac{1 \times 384}{8 \times 99}$, is the contribution of this particular cell to the total S . In the same way, the contribution to S of the next cell in this column is found to be $\frac{5^2 \times 384}{8 \times 25}$, and of the third and last cell,

TABLE 59

TO ILLUSTRATE THE CALCULATION OF C BY SHORTER METHOD
Boys: Ages 4-5 Years

		Weight in Pounds						
		24-28	29-33	34-38	39-43	44-48	49-53	Total
Height in Inches	45-47			1		2		3
	42-44			4	35	21	5	65
	39-41		5	87	90	7	1	190
	36-38	<u>1</u>	18	72	8			<u>99</u>
	33-35	5	15	5				25
	30-32	2						2

$$\begin{array}{lcl}
 \begin{array}{cccccc} 8 & 38 & 169 & 133 & 30 & 6 & 384 \end{array} \\
 \text{Column 1: } \frac{1}{8} \left[\frac{1}{99} + \frac{25}{25} + \frac{4}{2} \right] & = & .3762 \\
 \text{Column 2: } \frac{1}{38} \left[\frac{25}{190} + \frac{324}{99} + \frac{225}{25} \right] & = & .3264 \\
 \text{Column 3: } \frac{1}{169} \left[\frac{1}{3} + \frac{16}{65} + \frac{7569}{190} + \frac{5184}{99} + \frac{25}{25} \right] & = & .5549 \\
 \text{Column 4: } \frac{1}{133} \left[\frac{1225}{65} + \frac{8100}{190} + \frac{64}{99} \right] & = & .4671 \\
 \text{Column 5: } \frac{1}{30} \left[\frac{4}{3} + \frac{441}{65} + \frac{49}{190} \right] & = & .2792 \\
 \text{Column 6: } \frac{1}{6} \left[\frac{25}{65} + \frac{1}{190} \right] & = & .0650 \\
 & P = & \overline{2.0688}
 \end{array}$$

$$C = \sqrt{\frac{P-1}{P}} = \sqrt{\frac{1.0688}{2.0688}} = .72$$

$\frac{2^2 \times 384}{8 \times 2}$. These contributions from column 1 may be com-

bined to give $\frac{384}{8} \left(\frac{1}{99} + \frac{25}{25} + \frac{4}{2} \right)$. The contribution of each of the other five columns to S may be found in like manner. One further simplification can be made. Since N (i.e., 384) is a common factor in each column, it may be left out of the computations entirely, in calculating the contribution of each cell,

as shown in Table 59. Then if the sum of all six columns is denoted by P ,

$$C = \sqrt{\frac{P-1}{P}} \quad (105)$$

(short method of calculating C)

In the present case, C equals .72 and the coefficient of correlation, r , from the same table is .71 (see p. 286). The correspondence of C and r here is very close, closer perhaps than that generally secured, although the difference between the two coefficients is never very great when the conditions prescribed on page 391 are met. In the present case, N is large, the classification is 6×6 fold, and the distributions are fairly normal.

It may be of interest to compute χ^2 for this table. Since

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}, \quad \chi^2 = \frac{NC^2}{1 - C^2}, \text{ and in the present problem}$$

$\chi^2 = \frac{384 \times .52}{1 - .52} = 416$. Since $n = (6 - 1)(6 - 1)$ or 25, from Table 52 it is seen that when $\chi^2 = 416$ and $n = 25$, P lies far beyond the bounds of our table. This result is analogous to a D/σ_D of 25 or 30, say, i.e., it indicates an association which could not conceivably have arisen by chance, //

III. CURVILINEAR OR NON-LINEAR RELATIONSHIP

1. The Correlation Ratio

The relationship between the paired values of two sets of measures, X and Y , may be described in a general way as "linear" or "non-linear." When the means of the arrays of the successive columns and rows in a correlation table follow straight lines (at least approximately), the regression is said to be linear or straight-line (p. 291). When, however, the drift or trend of the means of the arrays (columns or rows) cannot be described by a straight line, but can be represented by a *curve* of some kind, the regression is said to be curvilinear or in general non-linear.

Our discussion in Chapter IX was concerned entirely with

linear relationship, the extent or degree of which is measured by the product-moment coefficient of correlation, r . It sometimes happens in mental measurement, however, that the relationship between two variables is definitely non-linear; and when this is true r is not an adequate measure of the degree of correspondence or correlation. When the regression is non-linear, the curve joining the means of successive arrays (in the columns, say) will fit these mean values more exactly than will a straight line. Hence, should truly curvilinear relationship be described by a straight line the scatter or spread of the paired values about the regression line will be greater than the scatter about the better-fitting regression curve. The smaller the spread of the paired scores about the regression line or the regression curve which relates the variables X and Y (or Y and X), the higher the relationship between the two variables. For this reason, an r calculated from a correlation table in which the regression is definitely curvilinear will *always be less* than the true relationship. An example will make the reason for this clearer. The correlation between the following two short series, as given by the product-moment formula, is $r = .93$ [formula (51), p. 271]. The *true* correlation between the two

Variable X	Variable Y
1	.25
2	.50
3	1.00
4	2.00
5	4.00

series, however, is clearly perfect, since changes in Y are directly related to changes in X . As X increases by 1 (i.e., in arithmetic progression) Y doubles (i.e., increases in geometric progression). The reason why r is less than 1.00 is obvious as soon as we plot the paired X and Y values. As shown in Figure 58, the relationship between X and Y is curvilinear, and is exactly described by a curve which passes through the successive plotted points. When linear relationship is forced upon these data, the plotted points do not fall along the straight

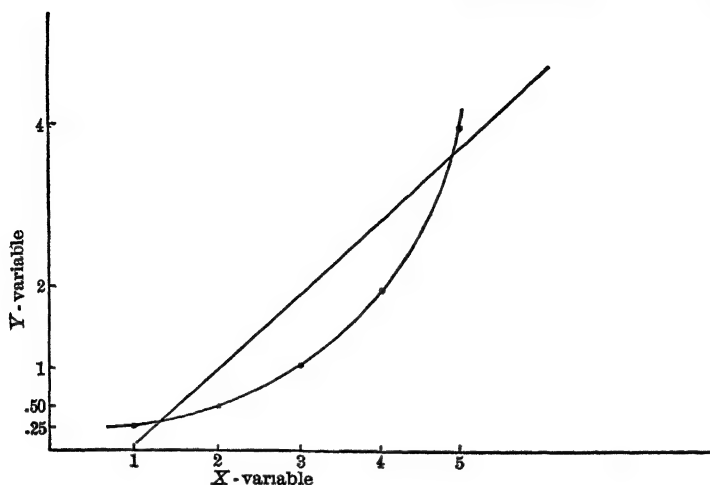


FIG. 58.

line, and the product-moment coefficient, r , is less than 1.00. However, the correlation-ratio, or coefficient of non-linear relationship η (read as *eta*) for the given data is 1.00.

Eta measures the concentration of paired X - and Y -values about a relation *curve* just as r measures the concentration of paired values about a relation *line*. *Eta* is a more general coefficient than r , however, as it is applicable when regression is linear, as well as when it is non-linear. If the regression is linear and the means fall along straight lines, η will equal r . But if regression is non-linear and the means lie along a curve, η will be greater than r . The coefficient of correlation, therefore, is a limiting value of the more general coefficient, η , just as straight-line relationship is a limiting case of curvilinear relationship. There are always two η 's in every correlation table, just as there are always two regression coefficients, $r \frac{\sigma_y}{\sigma_x}$ and $r \frac{\sigma_x}{\sigma_y}$, in a table in which regression is linear. The first correlation-ratio, written η_{yx} , gives the regression of Y on X (Y is the dependent variable). The second correlation-ratio, written

η_{xy} , gives the regression of X on Y (X is the dependent variable) [compare with the two regression equations (p. 290) in a correlation table in which relationship is linear].

The correlation-ratio is always *positive*, its value lying between .00 and 1.00. Whether the direction of relationship given by η is positive, negative, or a varying one, therefore, must be determined by inspection of the correlation diagram.

2. The Calculation of η in a Correlation Table

One of the most useful methods of calculating the two η 's (η_{xy} and η_{yx}) in a correlation table in which the relationship is known (or suspected) to be non-linear is illustrated in Figure 59.* Ordinarily, one will wish to compare the two calculated η 's with the r obtained from the same data in order to determine whether regression is, or is not, significantly non-linear. For this reason, the computation of r is included in Figure 59 as part of the process of calculating the η 's. The steps to be followed in finding η_{yx} † may be outlined briefly as follows:

Step 1

Construct a correlation table as shown in Figure 59, and described on page 258. Calculate σ_y and σ_x using the Assumed Mean method (p. 49).

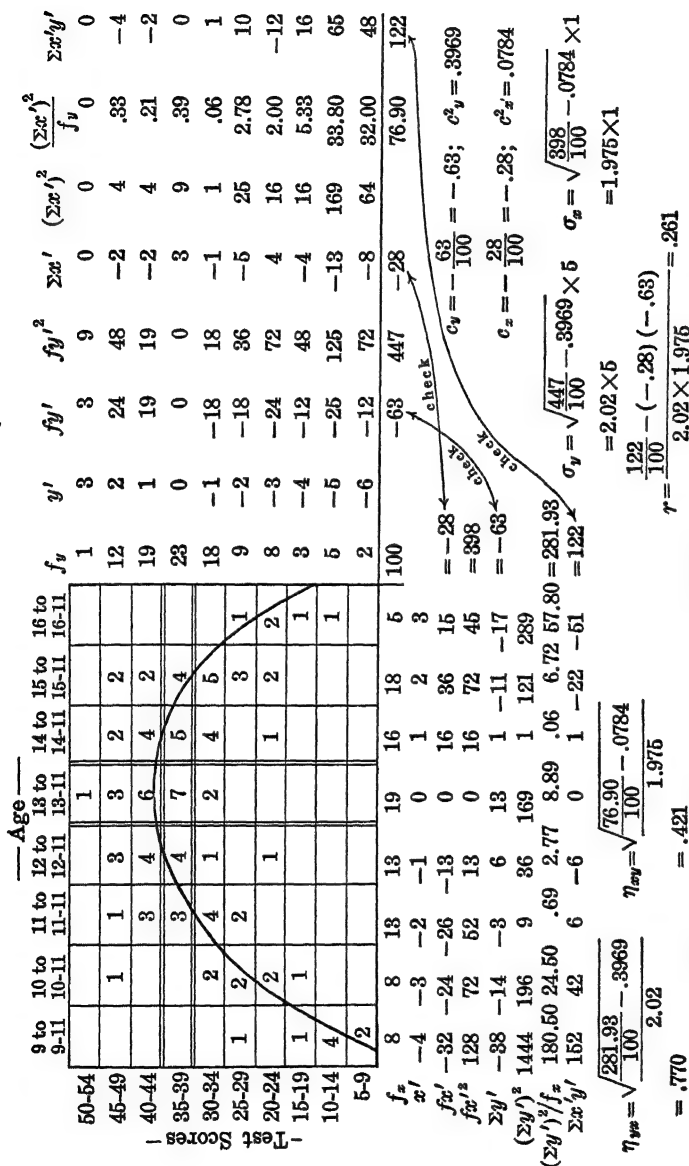
Step 2

Determine the entries in the $\Sigma y'$ column. These entries are found, as described on page 269, by multiplying the frequency in each column by its deviation (i.e., its y') measured in *units of step* from the Assumed Mean of the Y -distribution. To illustrate, in column 1, reading down, we have

* For further discussion of the method here outlined, see Dvorak, A., *A Simplified Computation of Non-linear Correlation*, Journal Educational Research, 1932, 25, pp. 99-104.

Holzinger, K. J., *A Combination Form for Calculating the Correlation Coefficient and Ratios*, Journal of the American Statistical Association, 1923, 18, pp. 623-627.

† The method of calculating η_{xy} , shown on the right of the diagram, follows exactly the method outlined here for the calculation of η_{yx} and hence will not be repeated.



$(1 \times -2) + (1 \times -4) + (4 \times -5) + (2 \times -6)$ or -38 . For column 2, the $\Sigma y'$ entry is $(1 \times 2) + (2 \times -1) + (2 \times -2) + (2 \times -3) + (1 \times -4)$ or -14 . Square each $(\Sigma y')$ entry to give the $(\Sigma y')^2$ row. Then divide each entry in the $(\Sigma y')^2$ row by its corresponding f_x to give the row $\frac{(\Sigma y')^2}{f_x}$. In column 1, for example, divide 1444 by 8 to obtain 180.50; and in column 2, divide 196 by 8 to obtain 24.50. The total of the $\frac{(\Sigma y')^2}{f_x}$ row in Figure 59 is 281.93.

Step 3

From $\frac{(\Sigma y')^2}{f_x}$, c^2_y , N , and σ_y , calculate η_{yx} by the following formula*:

$$\eta_{yx} = \frac{\sqrt{\frac{\frac{(\Sigma y')^2}{f_x}}{N} - c^2_y}}{\sigma_y} \quad (106)$$

(correlation-ratio, η_{yx} , a measure of non-linear relationship in terms of the standard deviation of the means of the Y-arrays)

In Figure 59 $\frac{(\Sigma y')^2}{f_x} = 281.93$; $N = 100$; $c^2_y = .3969$; and $\sigma_y = 2.02$ (in units of step-interval). Substituting these values in formula (106) we obtain .770 as the value of η_{yx} .

The formula for η_{xy} , the second *eta* in a correlation table, is

$$\eta_{xy} = \frac{\sqrt{\frac{\frac{(\Sigma x')^2}{f_y}}{N} - c^2_x}}{\sigma_x} \quad (107)$$

(correlation-ratio, η_{xy} , a measure of non-linear relationship in terms of the standard deviation of the means of the X-arrays)

* There are several alternate formulas, equivalent to formula (106), which may be used in calculating η_{yx} , when the coefficient of correlation is not wanted. See Peters, C. C., and VanVoorhis, W. R., *Statistical Procedures and Their Mathematical Bases*, 1935, pp. 298-307.

Yule, G. U., *An Introduction to the Theory of Statistics* (9th ed.), 1929, pp. 204-207.

In the present problem, $\eta_{xy} = .421$ (see Fig. 59 for calculations). In most correlation tables, as illustrated here, the two η 's will differ in size, since their values depend, respectively, upon the scatter about the curve joining the means of the Y -arrays, and the scatter about the curve joining the means of the X -arrays. In any particular problem, also, one correlation-ratio will ordinarily be of greater interest than the other; just as in linear correlation, one regression equation is usually of greater interest than the other (p. 308). In Figure 59 η_{yx} is obviously more valuable than η_{xy} , since it gives the *change* in score (Y) resulting from *changes* in age (X) — Y is the dependent variable, and X is the independent variable (p. 294). The curve which describes the relation between age and score — the curve through the means of the Y -columns — has been sketched in on the correlation diagram. Note that this curve begins and ends low, reaching its peak in the middle of the age range. Both younger and older children in the grade make low scores, the highest scores being achieved by children in the middle of the age range. A probable reason for the obtained non-linear relationship between age and score is that the given test contains elements unfamiliar to, or inadequately learned by, the younger children, and items too difficult for the older (and probably duller) children. The best scores, therefore, are achieved by those in the middle of the age range. The product-moment r in Figure 59 is .261. The method of testing whether regression is significantly non-linear (by comparing the η 's and r) will be given on page 400.

3. The Probable Error of η

The probable error of a correlation-ratio may be calculated by the formula

$$PE_{\eta} = \frac{.6745(1 - \eta^2)}{\sqrt{N}} \quad (108)$$

(probable error of a correlation-ratio, η)

The PE of the η_{xy} of .421 is .055, and of the η_{yx} of .770 is .027 ($N = 100$). Both of these coefficients are more than four times their PE 's and hence are significant.

4. The Correction of an Obtained η

The size of an obtained η depends directly upon the fineness of grouping in the X - and Y -variables, as well as upon the size of N . When N is comparatively small, or when the number of arrays in X or Y is large, Pearson * has provided a correction for fine grouping which should be applied to the obtained η . The formula for a corrected η is

$$\text{Corrected } \eta = \sqrt{\frac{\eta_{obt}^2 - \frac{(\kappa - 3)}{N}}{1 - \frac{(\kappa - 3)}{N}}} \quad (109)$$

(correction of η for fineness of grouping)

in which κ equals the number of arrays in X or Y . To illustrate if we apply this correction to η_{yx} obtained from Figure 59, we have, upon substituting .770 for η_{yx} , 8 for the number of Y -arrays (i.e. columns), and 100 for N ,

$$\begin{aligned} \text{Corrected } \eta_{yx} &= \sqrt{\frac{(.770)^2 - .05}{1 - .05}} \\ &= .756 \end{aligned}$$

The correction here is small, since η_{yx} is large, and the number of Y -arrays fairly small. The correction which must be applied to η_{xy} is much larger. Thus, substituting .421 for η_{xy} , and 10 for the number of X -arrays (i.e. rows), we have

$$\begin{aligned} \text{Corrected } \eta_{xy} &= \sqrt{\frac{(.421)^2 - .07}{1 - .07}} \\ &= .340 \end{aligned}$$

When η is small, and the grouping fine (i.e., data classified into many step-intervals) the correction given by formula (109) may be considerable, and hence should be made.

5. Tests for Linearity of Regression

It is by no means always easy to tell from the appearance of a correlation table whether regression is linear or non-linear. It seems clear in Figure 59, from the curve joining the means of

* Pearson, Karl, *On the Correction Necessary for the Correlation-Ratio* η , *Biometrika*, 1923, 14, pp. 412-417.

the columns, that the regression of Y on X , at least, is non-linear. Further evidence of non-linearity is given by the fact that the coefficient of correlation calculated from Figure 59 is .261, very much smaller than the η_{yz} of .770. As stated on page 395, when regression is strictly linear $\eta = r$; and the greater the departure of the regression from linearity, in general the greater the discrepancy between η and r . The usual test of linearity* is that ζ (read as *zeta*), which equals $\eta^2 - r^2$, shall differ from zero by an amount which is not greater than that which might arise from fluctuations due to random sampling. To make this test, we must calculate in addition to ζ , PE_{ζ} . This PE is given by the formula

$$PE_{\zeta} = .6745 \times 2\sqrt{\zeta/N} \cdot \{\sqrt{(1-\eta^2)^2 - (1-r^2)^2} + 1\} \quad (110)$$

(probable error of ζ , i.e., $\eta^2 - r^2$)

The second radical in this formula is approximately 1.00. Hence, unless great accuracy is demanded by the problem, we may write formula (110) as

$$PE_{\zeta} = .6745 \times 2\sqrt{\zeta/N} \quad (111)$$

(approximation to formula (110))

In our problem (see Fig. 59) $\eta_{yz} = .770$ and $r = .261$. Accordingly, $\zeta = (.770)^2 - (.261)^2$ or .525; and from formula (111), $PE_{\zeta} = .098$. *Zeta*, therefore, is more than 5 times its PE ($\zeta/PE_{\zeta} = \frac{.525}{.098} = 5.4$) and we may be assured that the

regression of Y on X is significantly non-linear. (Table 35, p. 214, may be used conveniently to evaluate ζ/PE_{ζ} .)

The test for linearity of regression may also be applied to η_{xy} . Since $\eta_{xy} = .421$, $\zeta = (.421)^2 - (.261)^2$ or .109. From formula (111), $PE_{\zeta} = .045$ and $\zeta/PE_{\zeta} = 2.42$. Although not quite significant, the chances are 95 in 100 that the regression of X on Y is also non-linear (p. 214). It would probably be advisable, therefore, to consider both regressions as non-linear, especially since the uncorrected η is considerably larger than r .

* Blakeman, J., *On Tests for Linearity of Regression in Frequency Distributions*, *Biometrika*, 1905, 4, pp. 332-350

When both η and r are small so that relationship is slight in any case, a simpler test for linearity (called "Blakeman's test") which does not require the calculation of a PE is often employed. According to this test when

$$N(\eta^2 - r^2) < 11.37 \quad (112)$$

the regression is linear.* Applying this test to the problem in Figure 59 we find that $N(\eta^2_{yx} - r^2) = 52.48$, clearly indicating non-linearity of regression. For η_{xy} , $N(\eta^2_{xy} - r^2) = 10.91$. This result indicates, as did the *zeta* test above, that the regression *may* be treated as linear. However, the odds are in favor of a better fit if a curve instead of a straight line is used.†

True non-linear relationship is often encountered in psychophysics and in experiments dealing with fatigue, practice, forgetting and learning. Whenever an experiment is carried on over a long period of time, so that results exhibit diminishing returns, relationship will be curvilinear. Most mental and educational tests, however, show linear or approximately linear relationships; and for this reason, r has been employed in psychology and education to a far greater extent than has η . If regression is significantly non-linear it makes considerable difference whether η or r is the measure of relation. However, if the correlation is low, and the regression not significantly curvilinear, r will give as adequate a measure of relationship as will η .

The coefficient of correlation has a distinct advantage over η in that knowing r we can write down at once the straight-line regression equation connecting X and Y or Y and X . This is not possible with the correlation ratio. In order to estimate one variable from another (say, Y from X) when regression is curvilinear, a curve must be fitted to the means of the Y -columns. The equation of this curve then serves as a "regression equation" from which estimates are made.‡

* The symbol $<$ means "less than"

† For criticism of Blakeman's test, see Fisher, R. A., *Statistical Methods for Research Workers*, 1930, p. 225.

‡ For an elementary treatment of curve fitting, see Peters, C. C., and VanVoorhis, W. R., *Statistical Procedures and Their Mathematical Bases*, 1935, pp. 308-323.

PROBLEMS

1. Compute the correlation between the following two series of test scores by
 - (a) the rank-difference method, and by
 - (b) the method of gains.

Individual	Intelligence Score (Army Alpha)	Cancellation Score (A-Test + Number Group Check- ing Test)
1	185	110
2	203	98
3	188	118
4	195	104
5	176	112
6	174	124
7	158	119
8	197	95
9	176	94
10	138	97
11	126	110
12	160	94
13	151	126
14	185	120
15	185	118

(Note: The cancellation scores are in *seconds*; hence the smallest score numerically (i.e., 94) is highest and is ranked No. 1.)

2. Check the product-moment correlations obtained in problems 5 and 7, pages 287-288, Chap. IX, by the rank-difference method.
3. The following data give the distributions of scores on the Thorndike Intelligence Examination made by entering college freshmen who presented *more* than 12 recommended units, and entering freshmen who presented *less* than 12 recommended units. Compute bi-serial r and its PE .

Thorndike Scores	More than 12 recommended units	Less than 12 recommended units
90-99	6	0
80-89	19	3
70-79	31	5
60-69	58	17
50-59	40	30
40-49	18	14
30-39	9	7
20-29	5	4
	<u>186</u>	<u>80</u>

4. The following data give the distributions of scores on Army Alpha made by those who answered 50% or more, and those who answered less than 50% of the items in test 2 (arithmetic) correctly.

Army Alpha Scores	Subjects answering 50% or more of the items on test 2 correctly	Subjects answering less than 50% of the items on test 2 correctly
185-189	7	0
175-184	16	0
165-174	10	6
155-164	35	15
145-154	24	40
135-144	15	26
125-134	10	13
115-124	3	5
105-114	0	5
	<u>120</u>	<u>110</u>

Compute bi-serial r and its PE .

5. Compute the tetrachoric r 's and their PE 's for the following tables which show
- The resemblance of grandchildren and grandparents with respect to eye color.
 - The correspondence of Yes and No answers to two items of a neurotic inventory.

Grandparents' Eye Color	Children's Eye Color		Totals	
	Light	Dark		
	Dark	185	203	388
	Light	450	275	725
Totals	635	478	1113	

Question 2	B. Question 1		Totals	
	No	Yes		
	Yes	83	187	270
	No	102	93	195
Totals	185	280	465	

6. (a) In 120 throws of a single die the following distribution of faces was obtained:

	Faces					
	1	2	3	4	5	6
Observed frequencies:	30	25	18	10	22	15
Total	120					

Do these results represent a significant deviation from expectation ("chance")? (Calculate χ^2 [p. 378] and test by Table 52.)

- (b) The following represents the number of boys and the number of girls, in a group of 160, who chose each one of the five alternative answers to an item in an attitude scale:

	Strongly Approve	Approve	Indifferent	Disapprove	Strongly Disapprove	Total
Boys	25	30	10	25	10	100
Girls	10	15	5	15	15	60

Do these data indicate a significant sex difference in attitude toward this question (see p. 386)?

7. Calculate the coefficient of contingency, C , for the three tables given below; also compute χ^2 for each table, and compare with C . Are the two findings in agreement?

A. Athletic Capacity—First Brother		Athletic	Betwixt	Non-athletic	Totals
Second Brother	Athletic				
	Betwixt	906	20	140	1066
	Non-athletic	20	76	9	105
	Totals	140	9	370	519
		1066	105	519	1690

Average amount of time spent in daily preparation	B.	Class In High School English							Totals
		9B	9A	10B	10A	11B	11A	12B	
More than one hour		31	5	9	0	0	0	3	48
31-60 mins.		81	24	75	28	66	19	4	297
16-30 mins.		51	13	36	10	15	28	8	161
1-15 mins.		9	6	7	2	3	1	1	29
No time		1	0	0	0	0	1	0	2
Totals		173	48	127	40	84	49	16	537

Education	C.	Salary						Totals
		0-900	901-1200	1201-2000	2001-4000	4001-10,000	10,001-	
Post Graduate Work					4	1		5
College Graduate				1	30	5	1	37
Business College			1	15	6	1		23
High School		2	10	30	7	1		50
Junior High		7	42	27	3	1		80
Elementary School		19	48	4	1			72
Totals		28	101	77	51	9	1	267

8. The following table shows the relationship between scores upon the Thorndike Intelligence Examination and extra-curricular activities of 102 Columbia College students.

(a) Compute η_{yx} and η_{xy} , and the PE 's.

(b) Test for linearity of regression.

Extra-curricular activities (Y)	Thorndike Scores (X)										
	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100-104	
18-20					2	2					4
15-17				2		3	1				6
12-14				4		6	2		2		14
9-11		1	2		4	4	6	7	3		27
6-8	1			6	2	2	6	2	4	1	24
3-5	1		1	3	5	3		5	1	1	20
0-2		1		1			1	1	1	2	7
	2	2	3	16	13	20	16	15	11	4	102

9. In the following table (a) calculate the two η 's and (b) test for linearity of regression.

		Age in Months (X)								
Scores on Test Y		80- 89	90- 99	100- 109	110- 119	120- 129	130- 139	140- 149	150- 159	f_y
	75-79								10	10
	70-74								12	12
	65-69								18	18
	60-64							8	16	24
	55-59							10	8	18
	50-54							12		12
	45-49							14		14
	40-44							6		6
	35-39						8	6		14
	30-34						19	7		26
	25-29				2	2	22	5		31
	20-24			1	10	17	26			54
	15-19		2	4	8	15	12			41
	10-14	5	5	12	8	24	9			63
	5-9	9	8	16	16	9	9			67
	0-4	6	6	3	20	13	7			55
	f_x	20	21	36	64	80	112	68	64	465

ANSWERS

1. (a) $\rho = .19$; $r = .20$
 (b) $R = .09$; $r = .16$
3. $r_{\text{bis}} = .34$; $PE_{r_{\text{bis}}} = .05$
4. $r_{\text{bis}} = .47$, $PE_{r_{\text{bis}}} = .05$
5. A. $r_t = .23$; $PE_{r_t} = .03$ (approximately)
 B. $r_t = .33$; $PE_{r_t} = .04$ (approximately)
6. (a) Yes. $\chi^2 = 12.90$; $n = 5$, and P is almost .02.
 (b) No $\chi^2 = 7.03$, $n = 4$, and P lies between .20 and .10.
7. A. $C = .68$; $\chi^2 = 1440$. P cannot be read from Table 52, but is extremely small. The C and χ^2 agree in indicating an association which cannot conceivably be attributed to "chance."
 B. $C = .36$; $\chi^2 = 80$. P is less than .01. The association is significant.
 C. $C = .70$; $\chi^2 = 257$. P is less than .01 and the association is significant.
8. (a) $\eta_{yz} = .43 \pm .06$; $\eta_{yz}(\text{corrected}) = .35$
 $\eta_{xy} = .20 \pm .06$; $\eta_{xy}(\text{corrected}) = .00$
 (b) $r = -.09$. By Blakeman's test $N(\eta_{yz}^2 - r^2) = 17.34$; and the regression is non-linear. $N(\eta_{xy}^2 - r^2) = 3.26$, and the regression is linear, or at least not significantly non-linear.
9. (a) $\eta_{yz} = .931 \pm .005$; $\eta_{yz}(\text{corrected}) = .929$
 $\eta_{xy} = .818 \pm .010$; $\eta_{xy}(\text{corrected}) = .812$
 (b) $r_{xy} = .784$. By Blakeman's test $N(\eta_{yz}^2 - r^2) = 117.18$, and the regression is clearly non-linear. $N(\eta_{xy}^2 - r^2) = 25.11$, and the regression is again non-linear.

CHAPTER XIV

PARTIAL AND MULTIPLE CORRELATION

I. THE MEANING OF PARTIAL AND MULTIPLE CORRELATION

THE coefficient of correlation between test scores or other measures often represents the degree of relationship existing between the given variables plus, in addition, the indirect effect of other factors to which both variables are related. In measuring the correlation between two sets of scores, it is often desirable to eliminate or rule out if possible the influence of those factors which through their common relationship to the measures to be correlated either obscure results, or else make them difficult to interpret. To illustrate what is meant, suppose that the correlation between intelligence test scores and chronological age in a large group of children, whose ages range from 7 to 14 years, is .50; that the correlation between school achievement and chronological age in the same group is .40; and that the correlation between intelligence test scores and school achievement is .70. Now this last coefficient, namely, .70, is not simply a measure of the influence of intelligence upon school achievement; but is a measure of the influence of intelligence, *plus* the indirect effect of differences in age or maturity, upon school achievement. Since both intelligence test scores and school achievement tend to increase with chronological age, the correlation between these two measures (when chronological age is allowed to vary) will be raised owing to their common dependence upon degree of maturity.

In order to discover the relationship between intelligence and school achievement, uninfluenced by maturity, it is necessary to rule out the factor of age differences. This can be

accomplished *experimentally* by selecting children all of whom are of the same age. This procedure, however, offers difficulties, the principal one being that it is almost impossible to find a large sample of children all of whom are of *exactly* the same age. It becomes necessary, therefore, to determine what age range is permissible; and the more exactly we try to select children with respect to age, the smaller the number we shall have left in our group. In fact, the experimental control of a factor by the method of selection will often limit the size of the group so decidedly as to make the obtained correlations of little value.

Because of the difficulty which arises in attempting to control a factor or factors experimentally, a second method is often employed. This is the statistical method of *partial correlation*, by means of which the correlation between two variables may be computed with the influence of one or more other variables eliminated. To illustrate, if we wished to find the correlation between general intelligence and school achievement, uninfluenced by the factor of age differences, we should calculate the "partial correlation" between them, i.e., the correlation with age "partialled out." Such a partial coefficient gives the *net* correlation between general intelligence and school achievement for children of the *same age*; or the *net* correlation between intelligence and school achievement when age is a constant factor. Expressed in still another way, our partial coefficient tells us what relationship exists between general intelligence test scores and school achievement when differences in maturity no longer affect *either* variable.

A second illustration of partial correlation may be helpful. Suppose that a teacher finds in her class a high correlation between test scores in history and test scores in arithmetic reasoning. In looking for an explanation of this correlation (since there is apparently little reason to *expect* high relationship between these two abilities), she finds that achievement in arithmetic seems to depend in part, at least, upon ability to read and understand the problems. Obviously, ability to read

well is also a factor in determining achievement in history. Now by means of a third test, one of reading comprehension, suppose that our teacher determines the partial correlation between history achievement and arithmetic reasoning, that is, the correlation between these two variables when differences in reading comprehension have been rendered constant. If this partial coefficient is considerably smaller than the "whole" coefficient between history and arithmetic, the hypothesis that the apparent relationship was due largely to the common dependence of both tests upon reading is verified. When a factor (or factors) is "partialled out" from a given correlation the effect is to eliminate the differences among the individuals which are introduced by the variable thus controlled. In general, then, a coefficient of partial correlation may be said to represent in a convenient way the *net* relationship between two variables when the influence of one or more factors which might increase or decrease the relationship sought has been ruled out or held constant. The method of controlling factors through partial correlation may be employed whenever the correlation can be computed between the factor or factors to be controlled and the two variables the net correlation of which is desired. Since all of the data is utilized, partial correlation has a decided advantage over the experimental methods of control in many problems.

In addition to its value as a device for controlling conditions by eliminating the influence of "disturbing" or other factors, partial correlation is useful in other ways. For one thing it enables us to build up a regression equation involving three or more variables from which a criterion or other test score may be predicted when we know the scores made by a subject on the correlated tests. The accuracy of the regression equation in estimating criterion scores — its reliability as a prediction instrument — may be determined by the "multiple" coefficient of correlation. A multiple correlation coefficient gives the correlation between a single test and a team of tests. Or, expressed more accurately, the multiple coefficient of correlation

gives the relationship between the scores *actually obtained* on a test and the scores on the same test as *estimated* from the regression equation made up of the tests of the battery or team. The meaning and value of the multiple coefficient of correlation will be better understood when the student has worked through an actual problem such as that given in Table 60.

To summarize briefly, partial and multiple correlation may be thought of as representing an important extension of the theory and technique of simple or two-variable correlation to include problems which involve three or more variables.

II. AN ILLUSTRATIVE CORRELATION PROBLEM INVOLVING THREE VARIABLES

Perhaps the most straightforward approach to an understanding of the meaning of partial and multiple correlation, and of the techniques of calculation involved, is through the solution of a problem. The present section, therefore, will show the application of partial and multiple correlation to a three-variable problem. Following this, the general formulas and further applications of the method will be considered.

The problem in Table 60 is taken from a study by May* of the factors which influence "academic success." In that part of his study from which the present example is drawn, May wished to discover how accurately he could predict the academic success or scholastic achievement of 450 Syracuse freshmen from a knowledge of their general intelligence and of their study habits. Academic success was defined specifically as the number of credit or "honor" points obtained by a student at the end of his first semester in college. The number of honor points earned depended upon the number of *A*, *B*, and *C* grades made by the student in his freshman courses. A grade of *A* carried three honor points; a grade of *B* two honor points; a grade of *C* one honor point; and a grade of *D*, which was a

* May, M. A., *Predicting Academic Success*, Journal of Educational Psychology, 1923, 14, pp 429-440.

passing mark, carried no honor point credit. The maximum number of points which a freshman taking the regulation number of courses in one semester could obtain was 48.

General intelligence was measured by a combination of the Miller Mental Ability Test, and the Dartmouth Completion of Definitions Test. The first test contains 120 items and the second 40, so that the maximum score was 160. The scores of the 450 students ranged from 50 to 150, the distribution being fairly normal. As a measure of interest and application it was decided to take the average number of hours per week spent in study. Information with regard to study habits was obtained by means of a questionnaire given at the beginning and again at the middle of the first semester. Among other items in the questionnaire upon which information was requested were the number of hours spent per week at meals, in sleeping, etc. These and other questions were included, in order that the student might think that he was being checked upon the distribution of his total time, and not upon his study habits alone. The correlation between the student's two estimates of the number of hours which he spent in study (given on the first and second questionnaires) was .86, indicating a satisfactory degree of reliability.

As stated above, the main object of this study was to find how accurately the number of honor points which a student earns can be predicted from a knowledge of his study habits and his general intelligence. Other factors, of course, such as health, personality, previous preparation and the like, are undoubtedly of importance in determining the number of honor points received, as May indicates in his article. The two factors selected were chosen because they are not only important but are also objective and measurable. As the first step in solving the given problem, it will be necessary to calculate the partial coefficient which shows to what extent honor points are related to general intelligence when the variable factor of study hours per week is held constant. Also, the partial coefficient must be calculated which shows to what

extent honor points are related to study hours when the variable effect of general intelligence is rendered constant. Apart from the employment of these partial coefficients in the regression equation from which we predict honor points, the information which they yield will prove in itself to be of considerable interest. The solution of the problem is outlined in the following series of steps; the necessary data and calculations will be found in Table 60.

Step 1. The mean and σ of each series of measures and the intercorrelations are first calculated. These intercorrelations are product-moment r 's computed as shown in Chapter IX. The correlation between (1) honor points and (2) general intelligence, written r_{12} , is .60; the correlation between (1) honor points and (3) the number of hours spent on the average in study per week, written r_{13} , is .32; and the correlation between (2) general intelligence and (3) hours of study per week, written r_{23} , is $-.35$. The low correlation between honor points and study hours is of decided interest; but the most surprising correlation is the $-.35$ between study hours and general intelligence. Evidently the brighter the student, the less he studies.

Step 2. When we have calculated the intercorrelations of our three variables, the next step is to calculate the net correlation between (1) honor points and (2) general intelligence with the influence of (3) study hours partialled out or held constant. This net or partial coefficient of correlation, written $r_{12.3}$, is found from the following formula:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}} \quad \text{formula (113), p 422}$$

Substitution of the values of r_{12} , r_{13} and r_{23} in the formula gives a partial coefficient, $r_{12.3}$, of .80. This means that if *all* of our 450 students had studied exactly the same number of hours per week, the coefficient of correlation between honor points earned and general intelligence test scores would have been

TABLE 60

A CORRELATION PROBLEM INVOLVING THREE VARIABLES
(To Illustrate Partial and Multiple Correlation)

Step 1. Primary Data

(1) Honor Points	(2) General Intelligence	(3) Average Hours of Study per Week
$M_1 = 18.5$	$M_2 = 100.6$	$M_3 = 24$
$\sigma_1 = 11.2$	$\sigma_2 = 15.8$	$\sigma_3 = 6$
$r_{12} = .60$	$r_{13} = .32$	$r_{23} = -.35$

Step 2. Calculation of Partial Coefficients of Correlation

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} = \frac{.60 - .32(-.35)}{.9474 \times .9367} = .80 \quad (113)$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}} = \frac{.32 - .60(-.35)}{.8000 \times .9367} = .71 \quad (113)$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}} = \frac{-.35 - .60 \times .32}{.8000 \times .9474} = -.72 \quad (113)$$

Step 3. The Regression Equations, and Partial Regression Coefficients

$$\bar{x}_1 = b_{12.3}x_2 + b_{13.2}x_3 \quad (\text{Deviation Form}) \quad (115)$$

or $\bar{X}_1 = b_{12.3}X_2 + b_{13.2}X_3 + K \quad (\text{Score Form}) \quad (116)$

in which $b_{12.3} = r_{12.3} \frac{\sigma_1}{\sigma_2}$ and $b_{13.2} = r_{13.2} \frac{\sigma_1}{\sigma_3}$ (117)

Step 4. Calculation of the Partial σ 's

$$(1) \sigma_{1.23} = \sigma_1 \sqrt{1 - r_{12}^2} \sqrt{1 - r_{13}^2} = 11.2 \times .8000 \times .7042 = 6.3 \quad (114)$$

$$(2) \sigma_{2.13} = \sigma_2 \sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2} = 15.8 \times .9367 \times .6000 = 8.9 \quad (114)$$

$$(3) \sigma_{3.12} = \sigma_3 \sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2} = 6 \times .9367 \times .7042 = 4.0 \quad (114)$$

Step 5. Calculation of the Partial Regression Coefficients, and Partial Regression Equation

Substituting for $r_{12.3}$, $r_{13.2}$, $\sigma_{1.23}$, $\sigma_{2.13}$, $\sigma_{3.12}$, we have

$$b_{12.3} = .80 \times \frac{6.3}{8.9} = .57; \quad b_{13.2} = .71 \times \frac{6.3}{4.0} = 1.12$$

Hence the regression equation becomes:

$$\bar{x}_1 = .57x_2 + 1.12x_3 \quad (\text{Deviation Form})$$

or $\bar{X}_1 = .57X_2 + 1.12X_3 - 66 \quad (\text{Score Form})$

Step 6. Calculation of the Standard Error of Estimate

$$\sigma_{(\text{est } X_1)} = \sigma_{1.23} = 6.3 \quad (120)$$

$$PE_{(\text{est } Y_1)} = 6745 \times 6.3 = 4.2 \quad (121)$$

Step 7. Calculation of the Coefficient of Multiple Correlation

$$R_{1(23)} = \sqrt{1 - \frac{\sigma_{1.23}^2}{\sigma_1^2}} \quad (122)$$

$$= .83$$

.80 instead of .60, the obtained correlation coefficient. In other words, when each student spends the *same number of hours in study*, there is a closer correspondence between general intelligence test scores and honor points earned than there is when the number of study hours varies.

The partial coefficient of correlation between (1) honor points and (3) hours spent in study per week with (2) general intelligence partialled out, or its influence held constant, is found from the formula

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2}} \quad \text{formula (113), p. 422}$$

Substitution of the values of r_{13} , r_{12} and r_{23} gives a partial coefficient, $r_{13.2}$, of .71, as against an obtained coefficient (r_{13}) of .32. This result means that if our group possessed the same general intelligence * there would be a much closer correspondence between the number of honor points received and the number of hours spent in study than there is when the members of the group possess varying degrees of general intelligence. This is certainly the result to be expected.

The last partial coefficient of correlation $r_{23.1}$ equals $-.72$. This coefficient gives the net correlation between (2) general intelligence and (3) study hours when the influence of (1) honor points is held constant. It is found from the formula

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{13}^2}} \quad \text{formula (113), p. 422}$$

Like the two partial r 's above, we may interpret $r_{23.1}$ to mean that the correlation between general intelligence and hours spent in study in a group in which every student earns the same number of honor points would be much higher (in the *inverse* direction) than the "raw" correlation between the same two factors in a randomly selected group. By a randomly selected group is meant a group in which the number of honor points received by different students varies. It seems evident that

* By "same general intelligence" is meant the same *score* on the given general intelligence tests.

the brighter student not only studies less than the average and dull (since $r_{23} = -.35$) but that the brighter the student, the less he *needs* to study in order to reach a given standard of academic success — earn a given number of honor points.

Step 3. The partial coefficients of correlation having been calculated, we may now write the regression equation from which the most probable number of honor points which a student will receive may be estimated, given his score in general intelligence and the number of hours he studies per week. The regression equation for three variables (in *deviation form*) is written as follows:

$$\bar{x}_1 = b_{12.3}x_2 + b_{13.2}x_3 \quad \text{formula (115), p. 426}$$

In this equation x_1 which stands for honor points is the *dependent* variable, x_2 and x_3 which stand for general intelligence and study hours, respectively, are the *independent* variables. Note the resemblance of this equation to the simple regression equation for two variables $\bar{y} = b_{12} \cdot x$ (p. 292). If \bar{x}_1 is put for \bar{y} , and x_2 for x in this equation, we have $\bar{x}_1 = b_{12} \cdot x_2$. When written in *score form*, the regression equation for three variables becomes

$$(X_1 - M_1) = b_{12.3}(X_2 - M_2) + b_{13.2}(X_3 - M_3)$$

or transposing and collecting terms,

$$\bar{X}_1 = b_{12.3}X_2 + b_{13.2}X_3 + K \quad (\text{a constant}) \quad \text{formula (116), p. 426}$$

It is clear that before we can use this equation we must find the value of the *partial regression coefficients* $b_{12.3}$ and $b_{13.2}$. These may be found from the formulas

$$b_{12.3} = r_{12.3} \frac{\sigma_{1.23}}{\sigma_{2.13}} \quad \text{and} \quad b_{13.2} = r_{13.2} \frac{\sigma_{1.23}}{\sigma_{3.12}} \quad \text{formula (117), p. 426}$$

and as we already have the values of $r_{12.3}$ and $r_{13.2}$ it is only necessary that we find $\sigma_{1.23}$, $\sigma_{2.13}$, and $\sigma_{3.12}$ (the partial σ 's) in order to replace the partial regression coefficients in the equation by numerical values.

Step 4. The values of the partial σ 's may be found from the formulas *

$$\begin{aligned}\sigma_{1.23} &= \sigma_1 \sqrt{1 - r_{12}^2} \sqrt{1 - r_{13.2}^2} \\ \sigma_{2.13} &= \sigma_{2.31} = \sigma_2 \sqrt{1 - r_{23}^2} \sqrt{1 - r_{21.3}^2} \text{ formula (114), p. 423} \\ \sigma_{3.12} &= \sigma_{3.21} = \sigma_3 \sqrt{1 - r_{23}^2} \sqrt{1 - r_{13.2}^2}\end{aligned}$$

Substituting the known values of the raw and partial r 's in these formulas we find that $\sigma_{1.23} = 6.3$; $\sigma_{2.13} = 8.9$; and $\sigma_{3.12} = 4.0$. (For the calculations whereby these values are obtained, see Table 60.)

Step 5. From the partial σ 's and the partial r 's the numerical values of the partial regression coefficients $b_{12.3}$ and $b_{13.2}$ are found to be .57 and 1.12, respectively. We may now write the regression equation in deviation form as

$$\bar{x}_1 = .57x_2 + 1.12x_3$$

It is evident from this equation that scores on the general intelligence test and number of study hours per week must be weighted, respectively, as 1 : 2 ($1.12/.57 = 1.96$) when predicting honor points.

It may be noted at this point that while the partial coefficient of correlation $r_{23.1}$ is of interest as giving us the relation between general intelligence and hours spent in study for a constant number of honor points earned, this coefficient is not actually needed in the regression equation $\bar{x}_1 = b_{12.3}x_2 + b_{13.2}x_3$. In order to evaluate the constants $b_{12.3}$ and $b_{13.2}$ in our regression equation, we need *only* $r_{12.3}$ and $r_{13.2}$. In fact, in *any* problem involving three variables, only two partial coefficients of correlation need be computed, if we are interested simply in the prediction of X_1 scores from known values of X_2 and X_3 .

In order to write the regression equation for three variables in score form we replace x_1 by $(X_1 - 18.5)$; x_2 by $(X_2 - 100.6)$; and x_3 by $(X_3 - 24)$. The equation then becomes

$$\bar{X}_1 = .57X_2 + 1.12X_3 - 66$$

* See page 424.

Given a student's general intelligence test score (X_2) and the number of hours he spends in study per week (X_3), we may from this equation estimate the "most probable" number of honor points which he will receive during the first semester. To illustrate, suppose that a student has a general intelligence test score of 120 and that he studies on the average 20 hours per week; how many honor points will he most probably receive during the first semester? Substituting $X_2 = 120$ and $X_3 = 20$ in the regression equation, we find that

$$\bar{X}_1 = .57 \times 120 + 1.12 \times 20 - 66 = 25$$

The most probable number of honor points which this student will receive, therefore, using the given criteria as the basis of our estimate, is 25.

Step 6. This estimate, like every other "most probable" number of honor points predicted from the regression equation, has a certain "error of estimate." The standard error of estimate of *any* honor point prediction, that is, of any X_1 predicted from the regression equation, $\bar{X}_1 = b_{12.3}X_2 + b_{13.2}X_3 + K$ is written $\sigma_{(\text{est } X_1)}$, and equals $\sigma_{1.23}$ directly (p. 423). The $PE_{(\text{est } X_1)} = .6745 \times \sigma_{(\text{est } X_1)}$.

The standard error of estimate in the present problem is 6.3, and the $PE_{(\text{est } X_1)}$ is 4.2. In the illustration given above, therefore, the 25 estimated honor points have a $PE_{(\text{est } X_1)}$ of about 4 points. This means that the chances are even (50 in 100) that the given student will receive not less than 21 nor more than 29 honor points. The reliability of any given honor point estimate made from the regression equation may be found in exactly the same way.

Step 7. The final step in the solution of a three-variable correlation problem is usually the computation of the coefficient of multiple correlation. "Multiple r " is generally written R^* ; it has been defined (see p. 411) as the coefficient of correlation between the scores *actually made* on the given test and the

* Multiple R must not be confused with the R of the Spearman Foot-rule formula, page 363.

scores on the same test *predicted* from the regression equation. In the present problem, R gives the correlation between *earned* honor points (X_1) and honor points *estimated* by means of the two variables, general intelligence test scores (X_2) and hours of study (X_3), when these two are combined into a team by means of a regression equation. The formula for R when we are dealing with three variables is

$$R_{1(23)} = \sqrt{1 - \frac{\sigma_{1.23}^2}{\sigma_1^2}} \quad \text{formula (122), p. 430}$$

In the present problem $R_{1(23)} = .83$. This means that if the most probable number of honor points which each student in our group of 450 will receive is predicted from the regression equation given on page 418, the correlation between these 450 predicted scores and the 450 scores actually received will be .83. Multiple R tells us to what extent X_1 is determined by the combined action of X_2 and X_3 ; or, in the present instance, to what extent honor points are determined by general intelligence and number of study hours per week taken together

III. GENERAL FORMULAS FOR USE IN PARTIAL AND MULTIPLE CORRELATION

1. General Formulas for Partial r 's

We have found in Table 60 that in a correlation problem involving three variables, one is able by the method of partial correlation to find the net relationship between two variables when the influence of a third is ruled out or held constant. By an extension of the partial correlation method, we may obtain the net correlation between X_1 and X_2 when *two or more* variables have been held constant. The partial coefficient of correlation $r_{12.34}$, for example, means by analogy to $r_{12.3}$ that the correlation between X_1 and X_2 has been freed of the influence of *both* X_3 and X_4 ; and the partial coefficient of correlation $r_{12.34 \dots n}$ means that the correlation between X_1 and X_2 has been freed of the influence of a large number of disturbing factors.

TABLE 61

A TABLE TO INFER THE VALUE OF $\sqrt{1-r^2}$ FROM A GIVEN VALUE OF r

r	$\sqrt{1-r^2}$	r	$\sqrt{1-r^2}$	r	$\sqrt{1-r^2}$
.0000	1.0000	.3400	.9404	.6800	.7332
.01	.9999	.35	.9367	.69	.7238
.02	.9998	.36	.9330	.70	.7141
.03	.9995	.37	.9290	.71	.7042
.04	.9992	.38	.9250	.72	.6940
.05	.9987	.39	.9208	.73	.6834
.06	.9982	.40	.9165	.74	.6726
.07	.9975	.41	.9121	.75	.6614
.08	.9968	.42	.9075	.76	.6499
.09	.9959	.43	.9028	.77	.6380
.10	.9950	.44	.8980	.78	.6258
.11	.9939	.45	.8930	.79	.6131
.12	.9928	.46	.8879	.80	.6000
.13	.9915	.47	.8827	.81	.5864
.14	.9902	.48	.8773	.82	.5724
.15	.9887	.49	.8717	.83	.5578
.16	.9871	.50	.8660	.84	.5426
.17	.9854	.51	.8617	.85	.5268
.18	.9837	.52	.8542	.86	.5103
.19	.9818	.53	.8480	.87	.4931
.20	.9798	.54	.8417	.88	.4750
.21	.9777	.55	.8352	.89	.4560
.22	.9755	.56	.8285	.90	.4359
.23	.9732	.57	.8216	.91	.4146
.24	.9708	.58	.8146	.92	.3919
.25	.9682	.59	.8074	.93	.3676
.26	.9656	.60	.8000	.94	.3412
.27	.9629	.61	.7924	.95	.3122
.28	.9600	.62	.7846	.96	.2800
.29	.9570	.63	.7766	.97	.2431
.30	.9539	.64	.7684	.98	.1990
.31	.9507	.65	.7599	.99	.1411
.32	.9474	.66	.7513	1.00	.0000
.33	.9440	.67	.7424		

In every partial coefficient of correlation, e.g., $r_{12.34}$, the subscripts to the *left* of the point (1 and 2), called *primary* subscripts, denote the two variables whose net correlation we are seeking. The subscripts to the *right* of the point (3 and 4), called *secondary* subscripts, denote the variables ruled out or held constant. The *order* in which the secondary subscripts are written is immaterial, i.e., $r_{12.34} = r_{12.43}$. The order of the primary subscripts is of importance, however, as it tells us

which variable is taken to be dependent and which independent. Thus r_{12} means that X_1 is dependent — is to be predicted from X_2 ; while r_{21} means that X_2 is dependent — is to be predicted from X_1 . The numerical values r_{12} and r_{21} are, of course, the same. The order of a partial r is determined by the *number* of its secondary subscripts. Thus $r_{12.3}$ is a partial r of the *first* order; r_{12} (an "entire" or "total" r) is a coefficient of zero order; $r_{12.345}$ is a coefficient of the *third* order.

The general formula for a partial r of the n th order is

$$r_{12.34 \dots n} = \frac{r_{12.34 \dots (n-1)} - r_{1n.34 \dots (n-1)} r_{2n.34 \dots (n-1)}}{\sqrt{1 - r_{1n.34 \dots (n-1)}^2} \sqrt{1 - r_{2n.34 \dots (n-1)}^2}} \quad (113)$$

(partial correlation coefficient in terms of the coefficients
of lower order — n variables)

From this formula partial r 's of any given order may be found. In a five-variable problem, for example, $(n - 1) = 4$, and $n = 5$, so that $r_{12.345}$ is written

$$r_{12.345} = \frac{r_{12.34} - r_{15.34} r_{25.34}}{\sqrt{1 - r_{15.34}^2} \sqrt{1 - r_{25.34}^2}}$$

that is, in *terms* of the partial r 's of the *second* order. These second order partial r 's must then be computed by formula (113) from r 's of the *first* order before the third order r , $r_{12.345}$, can be evaluated. In order to find partial r 's of a higher order, we must, in every case, first express these in terms of partial r 's of the next lower order; and these r 's, in turn, in terms of r 's of the next lower order, and so on until r 's of zero order have been reached.* In other words, it is necessary to "work up" from zero order r 's, whenever r 's of any higher order are to be computed. With the addition of each new variable the calculation is greatly increased. As a result, unless the work is carefully planned, the arithmetic soon becomes extremely laborious.

The *PE* of a partial r of any order may be found, like the *PE* of an entire r , by substituting in formula (57) or reading the value from Table 41, page 280.

* In calculating partial r 's use Table 61 to read $\sqrt{1 - r^2}$ values.

There are other methods of eliminating the influence of a variable or variables which are useful in certain correlational problems. We may mention briefly two of these, namely, *part correlation* and *semi-partial correlation*. Both of these methods differ from partial correlation in that they give the net effect (upon the correlation between two tests) obtained by ruling out the influence of one or more test variables from only *one* of the two correlated measures, instead of from *both* (as in partial correlation). One may wish to know the correlation (semi-partial) between reaction time and speed of reading, say, when differences in the size of vocabulary are held constant with respect to the reading test *only*. Part correlation and semi-partial correlation have not been widely used in mental measurement; and their application is confined principally to special problems. For a discussion of formulas and illustrations see references below.*

2. General Formulas for Partial σ 's of any Order

Just as the correlation between two sets of scores can be determined when the influence of 1, 2, 3, . . . n factors is held constant, so the variability (σ) of any set of scores can be found when the influence of 1, 2, 3, . . . n factors is ruled out. As an illustration, let us consider $\sigma_{1.23}$ of Table 60. This partial σ gives the variability of X_1 (honor points) freed of the influence exerted by the two factors X_2 (general intelligence) and X_3 (study hours per week). The general formula for partial σ 's of any order is

$$\sigma_{1.234 \dots n} = \sigma_1 \sqrt{1 - r_{12}^2} \sqrt{1 - r_{13.2}^2} \sqrt{1 - r_{14.23}^2} \dots \sqrt{1 - r_{1n.23 \dots (n-1)}^2} \quad (114)$$

(partial σ for n variables)

This formula may be used to compute the net σ 's in those correlation problems which involve any number of variables.

* Ezekiel, M., *Methods of Correlation Analysis*, 1930, pp. 181-183.
Dunlap, J. W., and Cureton, E. E., *On the Analysis of Causation*,
Journal of Educational Psychology, 1930, 21, pp. 657-680.

In a five-variable problem, for example, $\sigma_{1.2345}$ is written

$$\sigma_{1.2345} = \sigma_1 \sqrt{1 - r_{12}^2} \sqrt{1 - r_{13.2}^2} \sqrt{1 - r_{14.23}^2} \sqrt{1 - r_{15.234}^2}$$

and by reference to formula (114) the other σ 's may be written:

$$\sigma_{2.1345} = \sigma_2 \sqrt{1 - r_{23}^2} \sqrt{1 - r_{24.13}^2} \sqrt{1 - r_{25.134}^2}$$

$$\sigma_{3.1245} = \sigma_3 \sqrt{1 - r_{34}^2} \sqrt{1 - r_{35.12}^2} \sqrt{1 - r_{36.124}^2}$$

$$\sigma_{4.1235} = \sigma_4 \sqrt{1 - r_{45}^2} \sqrt{1 - r_{46.123}^2} \sqrt{1 - r_{47.1235}^2}$$

$$\sigma_{5.1234} = \sigma_5 \sqrt{1 - r_{56}^2} \sqrt{1 - r_{57.1234}^2} \sqrt{1 - r_{58.12345}^2}$$

Each of these partial σ 's gives the variability remaining in a *single test* when the variability introduced by differences in score in the *other four tests* is ruled out or held constant. All of the above partial σ 's are of the *fourth* order since they have four secondary subscripts, and the order of a partial σ , like the order of a partial r , is determined by the number of its secondary subscripts.

By a simple rearrangement of the secondary subscripts, any higher order σ may be written in more than one way. A partial σ of the second order may be written in two ways: for example, $\sigma_{1.23}$ which is given on page 418 as

$$\sigma_{1.23} = \sigma_1 \sqrt{1 - r_{12}^2} \sqrt{1 - r_{13.2}^2}$$

may also be written

$$\sigma_{1.32} = \sigma_1 \sqrt{1 - r_{13}^2} \sqrt{1 - r_{12.3}^2}$$

In like manner $\sigma_{2.13}$ may be written

$$(1) \quad \sigma_{2.13} = \sigma_2 \sqrt{1 - r_{23}^2} \sqrt{1 - r_{21.3}^2}$$

or

$$(2) \quad \sigma_{2.31} = \sigma_2 \sqrt{1 - r_{23}^2} \sqrt{1 - r_{21.3}^2}$$

and $\sigma_{3.12}$ may be written

$$(1) \quad \sigma_{3.12} = \sigma_3 \sqrt{1 - r_{32}^2} \sqrt{1 - r_{31.2}^2}$$

or

$$(2) \quad \sigma_{3.21} = \sigma_3 \sqrt{1 - r_{32}^2} \sqrt{1 - r_{31.2}^2}$$

The alternate forms of a partial σ are useful as a check upon arithmetic calculations, and because they render unnecessary the calculation of otherwise unused and superfluous partial r 's. By using the *second* form of $\sigma_{2.13}$ and $\sigma_{3.12}$ instead of the *first* (see Table 60), for example, it becomes unnecessary to calculate $r_{23.1}$ so far as the calculation of the σ 's in the regression equation is concerned. Furthermore, if $r_{23.1}$ is not wanted elsewhere in the problem, it need not be calculated at all (see p. 418). Two partial r 's are all that are required in order to write the regression equation of a three-variable problem.

The number of alternate forms in which any higher order σ may be written depends upon the number of permutations which its secondary subscripts can take. We have seen that a second order σ may be written in two ways: $\sigma_{1.23}$ and $\sigma_{1.32}$. In the same way, any σ of the third order, e.g., $\sigma_{1.234}$, may be written in six ways: $\sigma_{1.234}$, $\sigma_{1.243}$, $\sigma_{1.324}$, $\sigma_{1.342}$, $\sigma_{1.423}$, $\sigma_{1.432}$. Any σ of the fourth order, e.g., $\sigma_{1.2345}$, may be written in 24 ways and any σ of the fifth order, e.g., $\sigma_{1.23456}$, in 120 ways.*

Fortunately, we need only a few of these many possible arrangements. Care, nevertheless, must be exercised that the correct forms are chosen, for just as the number of partial r 's which must be computed in a three-variable problem can be reduced by a judicious choice of σ -formulas, so also in problems which involve more than three variables the number of partial r 's may be reduced by proper selection. In the longer problems a reduction in the number of partial r 's counts most, as it is here that the calculations become laborious. The partial σ 's which require the computation of the minimum number of partial r 's will be found (for four- and five-variable problems) in the outline solutions on pages 433-437. These will be useful for reference.

* The permutations of 4 things taken 4 at a time are ${}_4P_4 = 4 \times 3 \times 2 \times 1 = 24$; and the permutations of 5 things taken 5 at a time are ${}_5P_5 = 5 \times 4 \times 3 \times 2 \times 1 = 120$. In general, permutations of n things taken n at a time are ${}_nP_n = n(n-1)(n-2) \dots$ to n factors. See the chapter on Permutations and Combinations in any algebra text.

3. General Formulas for the Regression Equations and Partial Coefficients of Regression

(1) The Regression Equation for Any Number of Variables

The general regression equation which expresses the relation between a single dependent variable, X_1 , and any number of independent variables, $X_2, X_3, X_4, \dots, X_n$ may be written in *deviation form* as follows:

$$\bar{x}_1 = b_{12.34 \dots n}x_2 + b_{13.24 \dots n}x_3 + \dots + b_{1n.23 \dots (n-1)}x_n \quad (115)$$

(*regression equation, deviation form, for n variables*)

and in *score form*

$$\bar{X}_1 = b_{12.34 \dots n}X_2 + b_{13.24 \dots n}X_3 + \dots + b_{1n.23 \dots (n-1)}X_n + K \quad (116)$$

(*regression equation, score form, for n variables*)

The partial regression coefficients $b_{12.34 \dots n}, b_{13.24 \dots n}$, etc., give the *weight* or value to be attached to the score of each independent variable when X_1 is to be estimated from all of these in combination. Furthermore, the regression coefficients give the weight which each test has in determining X_1 , when the influence of the other tests is excluded. Hence, we can tell from the regression equation just what rôle each of the several tests plays in determining the score on Test 1, the test taken as the dependent variable.

The partial regression coefficients in a regression equation may be computed from the formula

$$b_{12.34 \dots n} = r_{12.34 \dots n} \frac{\sigma_{1.234 \dots n}}{\sigma_{2.134 \dots n}} \quad (117)$$

(*partial regression coefficients in terms of partial coefficients of correlation, and standard errors of estimate — n variables*)

When the problem involves only three variables, the regression coefficients, $b_{12.3}$ and $b_{13.2}$ are, like $r_{12.3}$ and $r_{13.2}$, of the *first order*.

The first regression coefficient, $b_{12.3}$, equals $r_{12.3} \frac{\sigma_{1.23}}{\sigma_{2.13}}$ and the

second regression coefficient, $b_{13.2}$, equals $r_{13.2} \frac{\sigma_{1.23}}{\sigma_{3.12}}$. Regression

equations which involve more than three variables may be written by reference to formula (116); and their partial regression coefficients may be calculated from formula (117). In a five-variable problem, for example, the regression equation becomes

$$\bar{X}_1 = b_{12.345}X_2 + b_{13.245}X_3 + b_{14.235}X_4 + b_{15.234}X_5 + K$$

and the regression coefficients (of the *third* order) are

$$b_{12.345} = r_{12.345} \frac{\sigma_{1.2345}}{\sigma_{2.1345}}$$

$$b_{13.245} = r_{13.245} \frac{\sigma_{1.2345}}{\sigma_{3.1245}}$$

$$b_{14.235} = r_{14.235} \frac{\sigma_{1.2345}}{\sigma_{4.1235}}$$

$$b_{15.234} = r_{15.234} \frac{\sigma_{1.2345}}{\sigma_{5.1234}}$$

In order to compute these partial regression coefficients we must first compute the third order partial r 's, and the necessary partial σ 's. The calculation of the b 's is then a matter of substitution.

(2) The Regression Equation for Three Variables (Shorter Form)

In a problem which involves only three variables, the regression equation, as we have seen, is written

$$\bar{x}_1 = b_{12.3}x_2 + b_{13.2}x_3 \quad (\text{deviation form})$$

In this equation, the regression coefficients, $b_{12.3}$ and $b_{13.2}$, are of the *first* order. If the partial r 's and partial σ 's are of no especial interest, it is possible to compute these partial regression coefficients in a simpler way than from formula (117). If we replace second order σ 's by first order σ 's, our regression equation becomes

$$\bar{x}_1 = r_{12.3} \frac{\sigma_{1.3}}{\sigma_{2.3}} x_2 + r_{13.2} \frac{\sigma_{1.2}}{\sigma_{3.2}} x_3$$

or expanding $\sigma_{1.3}$, $\sigma_{2.3}$, $\sigma_{3.2}$ (by formula 114)

$$\bar{x}_1 = r_{12.3} \frac{\sigma_1 \sqrt{1 - r_{13}^2}}{\sigma_2 \sqrt{1 - r_{23}^2}} x_2 + r_{13.2} \frac{\sigma_1 \sqrt{1 - r_{12}^2}}{\sigma_3 \sqrt{1 - r_{23}^2}} x_3$$

If now we substitute in this equation for $r_{12.3}$ and $r_{13.2}$, and cancel the radicals $\sqrt{1 - r_{13}^2}$ and $\sqrt{1 - r_{12}^2}$ in numerator and denominator, we have

$$\bar{x}_1 = \frac{\sigma_1(r_{12} - r_{13}r_{23})}{\sigma_2(1 - r_{23}^2)} x_2 + \frac{\sigma_1(r_{13} - r_{12}r_{23})}{\sigma_3(1 - r_{23}^2)} x_3 \quad (118)$$

(regression equation for three variables, shorter form)

or in score form

$$\bar{X}_1 = \frac{\sigma_1(r_{12} - r_{13}r_{23})}{\sigma_2(1 - r_{23}^2)} X_2 + \frac{\sigma_1(r_{13} - r_{12}r_{23})}{\sigma_3(1 - r_{23}^2)} X_3 + K \quad (119)$$

(regression equation for three variables, shorter form)

This equation involves *only* zero order r 's and zero order σ 's; hence X_1 may be estimated from it without our having to compute any partial r 's or partial σ 's. We may illustrate the formula using the data given in Table 60, page 415. Substituting for $\sigma_1 = 11.2$, $\sigma_2 = 15.8$, $\sigma_3 = 6$, $r_{12} = .60$, $r_{13} = .32$, and $r_{23} = -.35$, we have

$$\bar{x}_1 = \frac{11.2(.60 + .32 \times .35)}{15.8(1 - .35^2)} x_2 + \frac{11.2(.32 + .60 \times .35)}{6(1 - .35^2)} x_3$$

$$\bar{x}_1 = .57x_2 + 1.12x_3$$

which checks the regression equation as calculated in Table 60.

To put this equation into score form we must substitute $(X_1 - 18.5)$ for x_1 , $(X_2 - 100.6)$ for x_2 , and $(X_3 - 24)$ for x_3 . This gives

$$\bar{X}_1 = .57X_2 + 1.12X_3 - 66$$

Formulas (118) and (119) are useful in calculating the coefficients to be substituted in a regression equation of a three-variable problem if, as pointed out above, the partial r 's and partial σ 's are not wanted. These formulas do not apply to problems which involve more than three variables.

4. General Formulas for Standard and Probable Errors of Estimate

All X_1 scores estimated from a regression equation have a standard error of estimate which measures the error made in taking scores estimated from a regression equation instead of *actual* scores (those earned on the test). The standard error of estimate may be found from the formula for $\sigma_{1.234 \dots n}$ as follows:

$$\sigma_{(\text{est } X_1)} = \sigma_{1.234 \dots n} \quad (120)$$

(standard error of estimate for n variables)

and the $PE_{(\text{est } X_1)}$ equals

$$PE_{(\text{est } X_1)} = .6745\sigma_{(\text{est } X_1)} \quad (121)$$

(probable error of estimate for n variables)

Since $\sigma_{1.234 \dots n}$ must be computed in order to evaluate the partial regression coefficients (p. 417), $\sigma_{(\text{est } X_1)}$ is always calculated in the course of the problem. The interpretation of the probable error of estimate has been made on page 419 from the data of Table 60. To repeat, we found in Table 60, that the $\sigma_{(\text{est } X_1)}$ of any estimated number of honor points is 6.3 and the $PE_{(\text{est } X_1)}$ is 4.2. Hence, the chances are even (50-50) that the "most probable" number of honor points predicted for any student by the regression equation will be in error by four points or less. We may feel sure that any given honor points estimate will not be in error by more than 4×4 or 16 points.

It is worth while examining somewhat further the meaning of the formula $\sigma_{(\text{est } X_1)}$. This standard error of estimate equals $\sigma_{1.234 \dots n}$; and the latter indicates the effect upon the variability of Test 1 obtained by eliminating (or holding constant) the influence of Tests 2, 3, 4, $\dots n$. The smaller $\sigma_{1.234 \dots n}$ is with respect to σ_1 , the greater the influence of the other tests upon Test 1's variability. Furthermore, the greater the degree to which the correlated tests (i.e., Tests 2, 3, 4, $\dots n$) are able to account for the variability of Test 1 the smaller $\sigma_{(\text{est } X_1)}$

will be and the more accurate the prediction of Test 1 scores from the regression equation. (See p. 443 for further discussion.)

The standard error of estimate is a *minimum* when the regression equation is used to estimate X_1 scores.* For this reason, values of X_1 which are predicted from the regression equation are the "best estimates" of the actual X_1 values which can be made by way of a linear equation containing the given variables. The regression equation $\bar{X}_1 = .57X_2 + 1.12X_3 - .66$ (see p. 418) will serve as an illustration of what this statement means. Since the correlation between X_1 and X_2 , X_1 and X_3 , and X_2 and X_3 is linear in every case, X_1 (honor points) can be estimated from this equation with a smaller error of estimate than from any other *linear* equation.

5. General Formula for R , the Coefficient of Multiple Correlation

The correlation between a single dependent variable X_1 and $(n - 1)$ independent variables combined by means of a regression equation is given by the formula

$$R_{1(23 \dots n)} = \sqrt{1 - \frac{\sigma_{1.23 \dots n}^2}{\sigma_1^2}} \quad (122)$$

(multiple correlation coefficient in terms of partial
 σ 's — n variables)

in which $R_{1(23 \dots n)}$ is the coefficient of multiple correlation, σ_1 is the standard deviation of the dependent series of scores (X_1) and $\sigma_{1.23 \dots n}$ gives the variability left in Test 1 when Tests 2, 3 . . . n are held constant through partial correlation. When there are only three variables, the multiple coefficient of correlation becomes

$$R_{1(23)} = \sqrt{1 - \frac{\sigma_{1.23}^2}{\sigma_1^2}}$$

when there are five variables

$$R_{1(2345)} = \sqrt{1 - \frac{\sigma_{1.2345}^2}{\sigma_1^2}}$$

* Yule, G. U., *An Introduction to the Theory of Statistics*, 1929, p. 231.

By analogy the R for six, seven, or any number of variables may be written by reference to formula (122).

The error of estimate is a minimum when the regression equation is employed in estimating X_1 scores (p. 430). From this fact it follows that the multiple coefficient of correlation gives the *maximum correlation* obtainable between actual X_1 scores and X_1 scores estimated from a knowledge of the independent variables $X_2, X_3 \dots X_n$ in the regression equation. The truth of this statement hinges upon the linearity of regression in *all* of the correlation tables. R is valuable in indicating how accurately a given combination of scores ("team of tests") represents the actual values of X_1 when these measures are combined in accordance with the "best" linear equation. R is always taken to be positive; hence, errors of sampling do not neutralize each other but tend to become cumulative. As a result, the PE of R — which is found from the formula for the PE of any product-moment r (p. 280) — is not an entirely adequate measure of the coefficient's reliability. In order to test the reliability of an obtained R , we should compare it with the value of that R which would be obtained from the same number of cases and the same number of variables if the variables were uncorrelated. The formula for this "chance R ," i.e., the R which would arise from fluctuations of sampling alone, is

$$R = \frac{\sqrt{n-1}}{\sqrt{N}} \quad (123)$$

*(multiple coefficient of correlation arising by chance
from n variables and N cases)*

in which n is the number of variables and N is the number of cases.* To illustrate this formula let us apply it to the three-variable problem in Table 60 in which n is equal to 3 and N is equal to 450. Substituting for N and n in the formula, we obtain an R equal to .07, which indicates a highly satisfactory degree of reliability for the obtained R of .83.

* Rosenow, Curt, *The Analysis of Mental Functions*, Psychological Monographs, 1917, 24, p. 20.

The square of multiple R gives the percentage of the variance* of X_1 (i.e., σ^2_1) attributable to the influence of the dependent variables $X_2, X_3, \dots X_n$. In the problem here considered $R^2 = .69$; hence 69% of what causes students to differ from one another in school achievement may be attributed to differences in general intelligence, as measured by the given tests, and to differences in study habits. The other 31% must be attributed to factors not considered in the given problem.

If we replace $\sigma_{1.23 \dots n}$ in formula (122) by its value in terms of the entire and partial r 's (see formula 114) we may write the general formula for $R_{1(234 \dots n)}$ as follows:

$$R_{1(234 \dots n)} = \sqrt{1 - [(1 - r^2_{12})(1 - r^2_{13.2}) \dots (1 - r^2_{1n.23 \dots (n-1)})]} \quad (124)$$

(multiple coefficient of correlation in terms of partial coefficients of correlation — n variables)

Since a higher order σ may be written in a variety of ways, the number depending upon its order (see p. 424), there are several alternate forms for R . These serve as valuable means of checking the accuracy of our arithmetical calculations. In a three-variable problem, for example, $R_{1(23)}$ may be written as

$$R_{1(23)} = \sqrt{1 - [(1 - r^2_{12})(1 - r^2_{13.2})]} \quad (124)$$

or as

$$R_{1(23)} = \sqrt{1 - [(1 - r^2_{13})(1 - r^2_{12.3})]}$$

and in a four-variable problem $R_{1(234)}$ may be calculated from

$$R_{1(234)} = \sqrt{1 - [(1 - r^2_{12})(1 - r^2_{13.2})(1 - r^2_{14.23})]}$$

and checked by

$$R_{1(342)} = \sqrt{1 - [(1 - r^2_{13})(1 - r^2_{14.3})(1 - r^2_{12.34})]}$$

* See pages 350-355.

IV. OUTLINE OF FORMULAS NEEDED IN CORRELATION PROBLEMS INVOLVING (A) FOUR VARIABLES, AND (B) FIVE VARIABLES

In a multiple correlation problem the chief task, usually, is to calculate the constants in the regression equation with a minimum of time and labor. When we are working with not more than three variables, the simplest plan is to write down the formula for the regression equation at once and deduce from it what partial r 's and higher order σ 's must be found in order to give us the partial regression coefficients. This method is illustrated in Table 60, although in the table the partial r 's were computed before the regression equation was written down. When there are more than three variables, the labor of calculating the regression coefficients becomes almost prohibitive unless the work is carefully mapped out. The following outlines give the formulas for calculating with the minimum amount of arithmetic the partial coefficients in a regression equation involving (A) four variables and (B) five variables. If the problem involves more than five variables the student should consult special methods for handling such problems given in the references below.*

(A) Formulas Needed in a Four-Variable Problem

(1) **Regression Equation.** The regression equation for four variables is written by reference to formula (116) as follows:

$$\bar{X}_1 = b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 + K$$

(2) **Regression Coefficients.** The three regression coefficients needed in (1) are found from formula (117) —

* Franzen, R., and Derryberry, M., *The Routine Computation of Partial and Multiple Correlation*, Journal of Educational Psychology, 1931, 22, pp. 641-651.

Griffin, H. D., *Simplified Schemas for Multiple Linear Correlation*, Journal of Experimental Education, 1933, 1, pp. 239-254.

Symonds, P. M., *Job Analysis Sheet for Computing Partial and Multiple Coefficients of Correlation and Regression Coefficients*, Teachers College Record, 1925, 27, pp. 52-69.

$$b_{12.34} = r_{12.34} \frac{\sigma_{1.234}}{\sigma_{2.134}}$$

$$b_{13.24} = r_{13.24} \frac{\sigma_{1.234}}{\sigma_{3.124}}$$

$$b_{14.23} = r_{14.23} \frac{\sigma_{1.234}}{\sigma_{4.123}}$$

These regression coefficients evidently require the computation of three second order partial r 's, and four third order σ 's.

(3) Partial r 's

To find

$$r_{12.34} = \frac{r_{12.3} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

To find

$$r_{13.24} = \frac{r_{13.2} - r_{12} r_{23}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2}}$$

To find

$$r_{14.23} = \frac{r_{14.2} - r_{12} r_{23}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2}}$$

we must find three first order partial r 's as follows:

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

$$r_{14.3} = \frac{r_{14} - r_{13} r_{24}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{24}^2}}$$

$$r_{24.3} = \frac{r_{24} - r_{23} r_{34}}{\sqrt{1 - r_{23}^2} \sqrt{1 - r_{34}^2}}$$

we must find three first order partial r 's as follows:

$$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2}}$$

$$r_{14.2} = \frac{r_{14} - r_{12} r_{24}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{24}^2}}$$

$$r_{34.2} = \frac{r_{34} - r_{23} r_{24}}{\sqrt{1 - r_{23}^2} \sqrt{1 - r_{24}^2}}$$

no partials of first order are needed other than those already found.

Note that a minimum of nine partial r 's must be computed, three of the second order and six of the first order. The nine first and second order r 's together with the six zero order r 's make 15 coefficients of correlation required in all.

(4) **Standard Deviations.** The four third order σ 's required may be found from the following formulas which make use of no partial r 's other than those already computed in (3) above. From formula (114):

$$\begin{aligned} \sigma_{1.234} &= \sigma_1 \sqrt{1 - r_{12}^2} \sqrt{1 - r_{13.2}^2} \sqrt{1 - r_{14.23}^2} \\ \sigma_{2.134} \text{ (i.e., } \sigma_{2.341}) &= \sigma_2 \sqrt{1 - r_{23}^2} \sqrt{1 - r_{24.3}^2} \sqrt{1 - r_{213.4}^2} \\ \sigma_{3.124} \text{ (i.e., } \sigma_{3.241}) &= \sigma_3 \sqrt{1 - r_{23}^2} \sqrt{1 - r_{34.2}^2} \sqrt{1 - r_{312.4}^2} \\ \sigma_{4.123} \text{ (i.e., } \sigma_{4.321}) &= \sigma_4 \sqrt{1 - r_{24}^2} \sqrt{1 - r_{24.3}^2} \sqrt{1 - r_{412.3}^2} \end{aligned}$$

The numerical values of the regression coefficients may now be computed and substituted in the regression equation.

(5) **Standard Error of Estimate**, $\sigma_{(\text{est } X_1)}$. From formulas (120) and (121) we find:

$$\begin{aligned}\sigma_{(\text{est } X_1)} &= \sigma_{1.234} \quad [\text{for value } \sigma_{1.234} \text{ see (4) above}] \\ PE_{(\text{est } X_1)} &= .6745 \sigma_{(\text{est } X_1)}\end{aligned}$$

(6) **Coefficient of Multiple Correlation, R** . In a four-variable problem the multiple coefficient, R , written $R_{1(234)}$, may be found from formula (122):

$$R_{1(234)} = \sqrt{1 - \frac{\sigma_{1.234}^2}{\sigma_1^2}}$$

$R_{1(234)}$ may also be written

$$R_{1(234)} = \sqrt{1 - [(1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2)]} \quad (124)$$

and checked by

$$R_{1(342)} = \sqrt{1 - [(1 - r_{13}^2)(1 - r_{14.3}^2)(1 - r_{12.34}^2)]}$$

(B) Formulas Needed in a Five-Variable Problem

(1) Regression Equation

$$\bar{X}_1 = b_{12.345}X_2 + b_{13.245}X_3 + b_{14.235}X_4 + b_{15.234}X_5 + K \quad (116)$$

(2) Regression Coefficients

$$\begin{aligned}b_{12.345} &= r_{12.345} \frac{\sigma_{1.2345}}{\sigma_{2.1345}} & b_{14.235} &= r_{14.235} \frac{\sigma_{1.2345}}{\sigma_{4.1235}} \\ b_{13.245} &= r_{13.245} \frac{\sigma_{1.2345}}{\sigma_{3.1245}} & b_{15.234} &= r_{15.234} \frac{\sigma_{1.2345}}{\sigma_{5.1234}}\end{aligned} \quad (117)$$

(3) **Partial r 's**. We must compute 22 partial r 's as follows (formula (113)):

(a)

To find $r_{12.345}$ write as $r_{12.453}$.
Then —

$$r_{12.453} = \frac{r_{12.45} - r_{13.45}r_{23.45}}{\sqrt{1 - r_{13.45}^2} \sqrt{1 - r_{23.45}^2}}$$

(b)

To find $r_{13.245}$ write as $r_{13.452}$.
Then —

$$r_{13.452} = \frac{r_{13.45} - r_{12.45}r_{23.45}}{\sqrt{1 - r_{12.45}^2} \sqrt{1 - r_{23.45}^2}}$$

To compute this r we need three partial r 's of the second order, viz. —

$$r_{12.45} = \frac{r_{12.4} - r_{15.4}r_{25.4}}{\sqrt{1 - r_{15.4}^2} \sqrt{1 - r_{25.4}^2}}$$

$$r_{13.45} = \frac{r_{13.4} - r_{15.4}r_{35.4}}{\sqrt{1 - r_{15.4}^2} \sqrt{1 - r_{35.4}^2}}$$

$$r_{23.45} = \frac{r_{23.4} - r_{25.4}r_{35.4}}{\sqrt{1 - r_{25.4}^2} \sqrt{1 - r_{35.4}^2}}$$

To compute these three r 's we need six r 's of the first order, viz. —

$r_{12.4}$	$r_{15.4}$	$r_{13.4}$
$r_{25.4}$	$r_{23.4}$	$r_{35.4}$

(c)

To find $r_{14.235}$ write without change —

$$r_{14.235} = \frac{r_{14.23} - r_{15.23}r_{45.23}}{\sqrt{1 - r_{15.23}^2} \sqrt{1 - r_{45.23}^2}}$$

To compute this r we need three partial r 's of the second order, viz. —

$$r_{14.23} = \frac{r_{14.2} - r_{13.2}r_{34.2}}{\sqrt{1 - r_{13.2}^2} \sqrt{1 - r_{34.2}^2}}$$

$$r_{15.23} = \frac{r_{15.2} - r_{13.2}r_{35.2}}{\sqrt{1 - r_{13.2}^2} \sqrt{1 - r_{35.2}^2}}$$

$$r_{45.23} = \frac{r_{45.2} - r_{34.2}r_{35.2}}{\sqrt{1 - r_{34.2}^2} \sqrt{1 - r_{35.2}^2}}$$

To compute these r 's we need six r 's of the first order, viz. —

$r_{14.2}$	$r_{13.2}$	$r_{15.2}$
$r_{34.2}$	$r_{35.2}$	$r_{45.2}$

To compute this r we need no partial r 's other than those already found in (a).

(d)

To find $r_{15.234}$ write without change—

$$r_{15.234} = \frac{r_{15.23} - r_{14.23}r_{45.23}}{\sqrt{1 - r_{14.23}^2} \sqrt{1 - r_{45.23}^2}}$$

To compute this r we need no partials other than those already found in (c).

Note that we must compute a minimum of four third order r 's, six second order r 's, and 12 first order r 's, 22 in all.

(4) **Standard Deviations.** The five fourth order σ 's required may be found from the following forms which make use of only those partial r 's already computed in (3):

$$\begin{aligned}\sigma_{1.2345} &= \sigma_1 \sqrt{1-r_{12}^2} \sqrt{1-r_{13.2}^2} \sqrt{1-r_{14.23}^2} \sqrt{1-r_{15.234}^2} \\ \sigma_{2.1345} \text{ (i.e., } \sigma_{2.4531}) &= \sigma_2 \sqrt{1-r_{24}^2} \sqrt{1-r_{25.4}^2} \sqrt{1-r_{23.45}^2} \sqrt{1-r_{212.345}^2} \\ \sigma_{3.1245} \text{ (i.e., } \sigma_{3.4521}) &= \sigma_3 \sqrt{1-r_{34}^2} \sqrt{1-r_{35.4}^2} \sqrt{1-r_{23.45}^2} \sqrt{1-r_{13.245}^2} \\ \sigma_{4.1235} \text{ (i.e., } \sigma_{4.2351}) &= \sigma_4 \sqrt{1-r_{24}^2} \sqrt{1-r_{34.2}^2} \sqrt{1-r_{45.23}^2} \sqrt{1-r_{14.235}^2} \\ \sigma_{5.1234} \text{ (i.e., } \sigma_{5.2341}) &= \sigma_5 \sqrt{1-r_{25}^2} \sqrt{1-r_{35.2}^2} \sqrt{1-r_{45.23}^2} \sqrt{1-r_{15.234}^2}\end{aligned}\quad (114)$$

(5) **Standard Error of Estimate, $\sigma_{(\text{est } X_1)}$**

$$\sigma_{(\text{est } X_1)} = \sigma_{1.2345} \quad [\text{see (4) above for value}] \quad (120)$$

$$PE_{(\text{est } X_1)} = 6745 \sigma_{(\text{est } X_1)} \quad (121)$$

(6) **Coefficient of Multiple Correlation, R**

$$R_{1(2345)} = \sqrt{1 - \frac{\sigma_{1.2345}^2}{\sigma_1^2}} \quad (122)$$

may be calculated by

$$R_{1(2345)} = \sqrt{1 - [(1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2)(1 - r_{15.234}^2)]} \quad (124)$$

and checked by

$$R_{1(4523)} = \sqrt{1 - [(1 - r_{14}^2)(1 - r_{15.4}^2)(1 - r_{12.45}^2)(1 - r_{13.245}^2)]}$$

V. A MULTIPLE CORRELATION PROBLEM INVOLVING FOUR VARIABLES

1. The Solution of a Four-Variable Problem

On page 419 we found that a student's honor points (X_1) could be estimated with a fair degree of accuracy from a knowledge of his general intelligence test score (X_2) and the number of hours he spends in study per week (X_3). The

TABLE 62. THE SOLUTION OF A CORRELATION PROBLEM INVOLVING FOUR VARIABLES

Primary Data:

(1) Honor Points

$$M_1 = 18.5$$

$$\sigma_1 = 11.2$$

(2) General Intelligence

$$M_2 = 100.6$$

$$\sigma_2 = 15.8$$

(3) Average Hours of Study per Week

$$M_3 = 24$$

$$\sigma_3 = 6$$

(4) Mean High School Grade

$$M_4 = 79$$

$$\sigma_4 = 7.5$$

$$r_{12} = .60$$

$$r_{13} = .32$$

$$r_{14} = .40$$

$$r_{23} = -.35$$

$$r_{24} = .36$$

$$r_{34} = .11$$

For scheme of solution from here on, see page 433.

(1) Regression Equation

$$\bar{X}_1 = b_{12 \cdot 34} X_2 + b_{13 \cdot 24} X_3 + b_{14 \cdot 23} X_4 + K$$

(116)

(2) Partial Regression Coefficients

$$b_{12 \cdot 34} = \frac{r_{12 \cdot 34}}{r_{23 \cdot 34}} = \frac{r_{12 \cdot 34}}{r_{23 \cdot 34}}$$

$$b_{14 \cdot 23} = \frac{\sigma_{1 \cdot 234}}{\sigma_{4 \cdot 23}} = \frac{\sigma_{1 \cdot 234}}{\sigma_{4 \cdot 23}}$$

(117)

(3) Partial r 's

To find

$$r_{12 \cdot 34} = \frac{r_{12 \cdot 3} - r_{14 \cdot 3} r_{24 \cdot 3}}{\sqrt{1 - r_{24 \cdot 3}^2} \sqrt{1 - r_{23 \cdot 3}^2}}$$

we must calculate three partial r 's of the first order:

To find

$$r_{13 \cdot 24} = \frac{r_{13 \cdot 2} - r_{14 \cdot 2} r_{24 \cdot 2}}{\sqrt{1 - r_{24 \cdot 2}^2} \sqrt{1 - r_{23 \cdot 2}^2}}$$

we must calculate three partial r 's of the first order:

To find

$$r_{14 \cdot 23} = \frac{r_{14 \cdot 2} - r_{13 \cdot 2} r_{24 \cdot 2}}{\sqrt{1 - r_{23 \cdot 2}^2} \sqrt{1 - r_{24 \cdot 2}^2}}$$

no first order partial r 's required other than the six already found.

(118)

$$r_{12 \cdot 3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

$$= \frac{.60 - .32 \times -.35}{.9474 \times .9367} = .80$$

$$r_{13 \cdot 2} = \frac{r_{13} - r_{13} r_{23}}{\sqrt{1 - r_{23}^2} \sqrt{1 - r_{23}^2}}$$

$$= \frac{.32 - .60 \times -.35}{.8 \times .9367} = .71$$

$$r_{14 \cdot 2} = \frac{r_{14} - r_{13} r_{24}}{\sqrt{1 - r_{23}^2} \sqrt{1 - r_{24}^2}}$$

$$= \frac{.40 - .32 \times .11}{.9474 \times .9939} = .39$$

$$r_{14 \cdot 23} = \frac{r_{14} - r_{13} r_{24}}{\sqrt{1 - r_{23}^2} \sqrt{1 - r_{24}^2}}$$

$$= \frac{.40 - .60 \times .36}{.8 \times .9330} = .25$$

$$r_{24} = \frac{r_{24} - r_{23}r_{34}}{\sqrt{1 - r_{23}^2} \sqrt{1 - r_{34}^2}} = .43$$

From the above

$$r_{12} = .80 - .39 \times .43 = .76$$

(4) Standard Deviations

$$\begin{aligned} \sigma_1 &= \sigma_1 \sqrt{1 - r_{12}^2} = 11.2 \times .8 \times 7042 \times .9959 = 6.3 \\ \sigma_2 &= \sigma_2 \sqrt{1 - r_{23}^2} = 15.8 \times .9367 \times 9028 \times .6499 = 8.7 \\ \sigma_3 &= \sigma_3 \sqrt{1 - r_{34}^2} = 6 \times .9367 \times .9629 \times .7238 = 3.9 \\ \sigma_4 &= \sigma_4 \sqrt{1 - r_{44}^2} = 7.5 \times .9939 \times .9028 \times .9959 = 6.7 \end{aligned}$$

$$b_{12} = r_{12} \frac{\sigma_1}{\sigma_2} = .76 \times \frac{6.3}{8.7} = .55$$

The computation of partial regression coefficients b_{12} , b_{13} , and b_{14} .

$$b_{13} = r_{13} \frac{\sigma_1}{\sigma_3} = .69 \times \frac{6.3}{3.9} = 1.12$$

$$b_{14} = r_{14} \frac{\sigma_1}{\sigma_4} = .09 \times \frac{6.3}{6.7} = .08$$

Substituting the values of the b 's, the regression equation becomes

$$\bar{x}_1 = .55x_2 + 1.12x_3 + .08x_4 \quad (\text{Deviation Form})$$

$$\bar{X}_1 = .55X_2 + 1.12X_3 + .08X_4 - 70 \quad (\text{Score Form})$$

$$\sigma_{(\text{est. } \bar{x}_1)} = \sigma_1 \sqrt{1 - r_{11}^2} = 6.3$$

$$PE_{(\text{est. } \bar{x}_1)} = 6745 \times 6.3 = 4.2$$

(5) Standard Error of Estimate

(6) Coefficient of Multiple Correlation R

$$R_{(123)} = \sqrt{1 - \frac{\sigma_{(\text{est. } \bar{x}_1)}^2}{\sigma_1^2}} = \sqrt{1 - \frac{(6.3)^2}{(11.2)^2}} = .83$$

Check;

$$R_{(234)} = \sqrt{1 - \frac{[1 - r_{12}^2](1 - r_{13}^2)(1 - r_{14}^2)]}{1 - [8976 \times .8479 \times 4224]}} = .83$$

$$r_{34} = \frac{r_{34} - r_{23}r_{24}}{\sqrt{1 - r_{23}^2} \sqrt{1 - r_{24}^2}} = .27$$

From the above

$$r_{13} = .71 - .25 \times .27 = .69$$

From known partial r 's

$$r_{14} = \frac{.25 - .71 \times .27}{.7042 \times .9629} = .09$$

(114)

(120)

(121)

(122)

(124)

$PE_{(\text{est } X_1)}$ made in estimating individual scores from the three-variable regression equation was 4.2; and the coefficient of multiple correlation, $R_{1(23 \dots n)}$, which indicates in general how well the estimated scores represent the actual scores, was .83. Now suppose we add to the two independent variables X_2 and X_3 a third variable X_4 , namely, the quality of preparatory work done by the student in high school.* We shall then have three independent variables from which to estimate the dependent variable, honor points, and the question arises, with how much greater accuracy will this additional variable enable us to predict academic success?

The answer to this question will be found in Table 62 which presents a solution of the problem following the scheme outlined for four-variable problems on page 433. Some additional discussion of procedure and methods, and several points to be noted especially, are given in the following paragraphs.

(1) Note that the mean and the σ of each of the four variables must be calculated as well as their six intercorrelations. With this primary data in hand, we may, following the outline on pages 433-435, write the regression equation and from it deduce what partial r 's and higher order σ 's will be required.

(2) It is clear from the coefficients in the regression equation that we shall need three partial r 's of the second order, $r_{12.34}$, $r_{13.24}$, and $r_{14.23}$; and four partial σ 's of the third order, $\sigma_{1.234}$, $\sigma_{2.134}$, $\sigma_{3.124}$, and $\sigma_{4.123}$, in order to evaluate the constants in the regression equation. Only those partial r 's actually required in the regression equation will be calculated.

(3) In order to find $r_{12.34}$, we shall need three first order partial r 's, $r_{12.3}$, $r_{14.3}$, and $r_{24.3}$; and to find $r_{13.24}$, we shall need three first order partial r 's, $r_{13.2}$, $r_{14.2}$, and $r_{34.2}$. For the last second order partial, $r_{14.23}$, no additional first order r 's are needed. A minimum of nine partial r 's, therefore, is required.

The partial $r_{12.34}$ gives the net correlation between (1) honor

* High school performance was measured by the average grade obtained in the work offered for college entrance. See May, M. A., *Predicting Academic Success*, Journal of Educational Psychology, 1923, 14, pp. 434-436.

points, and (2) general intelligence, when both (3) study hours and (4) average high school grades have been eliminated as variable factors, or their influence held constant. In like manner, $r_{13.24}$ gives the net correlation between (1) honor points and (3) study hours, when the variable influence of both (2) general intelligence and (4) average high school grades is held constant. The first second order partial r ($r_{12.34}$) equals .76 and is but slightly reduced from $r_{12.3}$ which equals .80; the second partial $r_{13.24}$ equals .69 and is also but slightly less than $r_{13.2}$ which equals .71. This comparison shows the relatively small influence of (4) high school grades upon the net correlation between (1) honor points and (3) study hours when the influence of (2) general intelligence is rendered constant; as well as the small influence of (4) high school grades upon the net correlation between (1) honor points and (2) general intelligence when the variable influence of (3) study hours is rendered constant. It is interesting to note that while the zero order coefficient of correlation between (1) honor points and (4) high school grades, viz., r_{14} , equals .40, $r_{14.2}$ equals .25, $r_{14.3}$ equals .39, and $r_{14.23}$ equals .09. Apparently, nearly all of the correlation which appears between (1) honor points and (4) high school grades may be attributed to the common dependence of these two variables upon (2) general intelligence and (to a somewhat lesser degree) upon (3) study hours.

(4) Following the outline on page 434 we are able to calculate the four third order σ 's required by the partial regression coefficients without the necessity of calculating any additional partial r 's. The partial σ 's, i.e., $\sigma_{1.234}$, $\sigma_{2.134}$, etc., give the net variability of the distribution of measures indicated by the *primary* subscript when the influence of *all three* of the other factors (denoted by the *secondary* subscripts) has been rendered constant. To illustrate, note that $\sigma_{1.234}$ is 6.3 while σ_1 is 11.2. This 44% reduction in variability means, concretely, that if *all* of the 450 students in the group were on a par as regards (2) general intelligence, (3) study hours, and (4) high school grades, the σ of their distribution of honor points would be

only about *one-half* as large as the observed σ — the σ of the group in which these factors are allowed to vary.

The computation of the partial regression coefficients is now simply a matter of combining the partial r 's and partial σ 's already found. When this has been done, we may substitute in the regression equation to find $\bar{x}_1 = .55x_2 + 1.12x_3 + .08x_4$. In score form the regression equation becomes $\bar{X}_1 = .55X_2 + 1.12X_3 + .08X_4 - 70$.

It is clear from the regression equation that (3) the number of hours spent in study has twice the weight of (2) the score on the general intelligence test and 14 times the weight of (4) the average high school grade in determining the number of honor points which a student will receive at the end of the first semester. The student must not be misled into thinking that these "weights" give the relative importance or "intrinsic" value of our three factors in determining academic success. The partial regression coefficients, to be sure, give the weights to be assigned to the *scores* made by the student in each of the independent variables. But the size of these weights depends directly upon the *units* in which the test is scored; hence they are essentially *conversion factors* which have the effect of changing the units in which the independent variables are expressed into the units of the dependent variable. The *relative importance* of each independent variable is determined by the method outlined on page 452

(5) Evidence of the small importance of high school grades in improving the estimate of honor points is given by $PE_{(est. X_1)}$. The PE of estimate made in predicting honor points from the present equation is 4.2 as compared with the $PE_{(est. X_1)}$ of 4.2 made in using the regression equation which does not include high school grades (p. 419). This probable error means that we can estimate the number of honor points that the student will receive, knowing his general intelligence score and the number of hours he studies per week, with no greater error than is made when we know in addition to these two the average grade received in high school. It is apparent, therefore,

that the work required to build up a regression equation which will include the factor of high school grades is not worth while.

(6) The multiple coefficient of correlation $R_{1(234)}$ is .83 as compared with the multiple $R_{1(23)}$ of .83. These multiple coefficients further substantiate the conclusion reached above on the basis of the $PE_{(est. X_1)}$ that high school grades contribute nothing to the precision of an honor point estimate.

It is of interest to compare the accuracy of our estimates of honor points, when the other three variables singly and in combination are taken into account. In this way the prognostic value of the multiple regression equation can be the more readily appreciated. The standard errors of estimate and the coefficients of correlation from the different variables taken singly and in combination are as follows:

Dependent variable (Honor points X_1)	$\sigma_{(est. X_1)}$	Coefficients of Correlation
$\bar{X}_1 = .60X_2 + 4.1$	10.6	$r_{12} = .32$
$\bar{X}_1 = .43X_3 + 24.8$	9.0	$r_{13} = .60$
$\bar{X}_1 = .57X_2 + 1.12X_3 - 66$	6.3	$R_{1(23)} = .83$
$\bar{X}_1 = .55X_2 + 1.12X_3 + .08X_4 - 70$	6.3	$R_{1(234)} = .83$

The important fact to be noted here is that $\sigma_{(est. X_1)}$ is considerably less, and the correlation considerably greater, when X_2 and X_3 are taken together than when either is taken alone. The standard error of estimate and the multiple R show no improvement when X_4 is added to X_2 and X_3 . It seems reasonable to expect, however, that by the addition of other variables to our regression equation, $\sigma_{(est. X_1)}$ could be still further reduced and R increased. But this is not immediately apparent without trial; and additional variables may, as shown above, contribute very little to our criterion estimates.

2. The Effect Upon Multiple R of Adding Tests to the Regression Equation

As noted in the section above, the further addition of tests to a battery does not always operate to increase appreciably the multiple correlation of the test battery with the criterion.

In many problems, in fact, "diminishing returns" appear after the addition of the third or fourth variable to the regression equation so that the inclusion of other variables in the test battery does not justify the time and calculation required. An example of "diminishing returns" was given on page 443 above, where it was seen that the addition of high school grades to the test battery (2) general intelligence and (3) hours of study had a negligible effect upon multiple R . Hull * has shown the net effect upon multiple R of including in a battery additional tests, all of which have correlations of a given size with each other, as well as correlations of the same size with the criterion. For example, a battery of five tests, each of which has a correlation of .40 with the criterion and of .20 with the others, will have an R of .667 with the criterion. The addition of 45 equivalent tests to the battery (making 50 in all) will raise the R to .860; but 50 *more* tests (making 100 in all) must be added to raise multiple R to .877 — .017 of a point!

Toops † has devised a method whereby tests may be added one after another to a battery so that the effect of each additional test upon multiple R can be estimated. The usefulness of this technique is illustrated in a study by Gates ‡ who found in a group of third grade children a correlation of .65 between educational achievement and a group intelligence test. Using Toops' method, Gates added Stanford-Binet M.A. to the group test, raising the correlation with achievement to .699, but adding a non-verbal battery of tests to these two only raised the correlation with achievement to .702 — a negligible gain. Flemming § obtained a correlation of .6339 between

* Hull, C. L., *The Joint Yield from Teams of Tests*, Journal of Educational Psychology, 1923, 14, pp. 396-406.

† Toops, H. A., *Tests for Vocational Guidance of Children Thirteen to Sixteen*, Teachers College, Columbia University, Contributions to Education, 1923, No. 136, pp. 137-153.

‡ Gates, A. I., *The Correlations of Achievement in School Subjects with Intelligence Tests and Other Variables*, Journal of Educational Psychology, 1922, 13, pp. 129-139, 223-235, 277-285.

§ Flemming, C. W., *A Detailed Analysis of Achievement in High School*, Teachers College, Columbia University, Contributions to Education, 1925, No. 196, p. 93.

Terman Group Test of Mental Ability and school achievement in a group of 60 third year high school students. Adding successively chronological age, the Miller Test of Mental Ability, the Otis Group Test, and the Haggerty Silent Reading to the Terman Test, she raised the multiple R with school achievement from .6339 to .6345 to .6349 to .6946 to .7015. Only the addition of the Otis Test produced a significant increase in the multiple R .

Before working out a regression equation containing an added variable or added variables, the predictive value of the new equation may be estimated by computing multiple R . This will enable one to determine whether correlation with the criterion is sufficiently increased to justify the additional calculation. When tests can be found which exhibit low correlations among themselves and high correlations with the criterion or dependent variable, they serve definitely to increase multiple R . Such tests contribute something not measured by the other members of the battery. It is difficult, however, to find distinctive tests of this sort, since tests which correlate highly with a common criterion usually correlate highly with each other. For this reason, a four- or five-variable test battery will often yield a multiple R which is almost — if not quite — as high as that obtained from a battery composed of many more tests.

VI. THE VALUE AND USE OF PARTIAL AND MULTIPLE CORRELATION

1. The Value of Partial Correlation in Description and Analysis

Partial correlation may be of decided value as an aid in analyzing the part played by each of several factors in determining a total result. An illustration may be cited from the work of Cyril Burt.* Burt wished to find to what extent a child's M.A., as measured by the Binet test, influences his school attainment. His subjects were 300 children, 7-14 years old.

(1) Each child's M.A. was determined; (2) his scholastic

* Burt, Cyril, *Mental and Scholastic Tests*, London, 1921, pp. 180-184.

achievement as measured by educational examinations and checked by teachers; and (3) his chronological age. The correlation between Binet M.A. and scholastic achievement (r_{12}) was .91. When chronological age (3) was partialled out the correlation ($r_{12.3}$) between Binet M.A. and scholastic achievement dropped to .68. This result shows, in the first place, that chronological age has a decided effect upon the correlation between M.A. and school work; it tends to increase or "dilate" the obtained r . This dilation is brought about by the fact that both M.A. and school attainment increase with C.A., and this common dependence on chronological age serves to boost the observed correlation. The residual partial correlation ($r_{12.3}$) of .68 indicates, however, a substantial relationship remaining between M.A. and school work when age is a constant factor. In other words, Binet M.A. is a substantial factor in a pupil's school attainment at each age level from 7 to 14. Taking the analysis a step further, Burt found that the correlation between (2) school work and (3) chronological age (r_{23}) was .87; and that when the effect of Binet M.A. was held constant, the partial r ($r_{23.1}$) between school work and C.A. was reduced to .49. This persistence of a substantial relationship between school work and C.A., when variability arising from differences in M.A. is eliminated, offers confirmatory evidence according to Burt of the "undue influence of age upon school classification." In these illustrations it is clear that the partial r 's constitute the first step in an analysis of the factors determining school achievement. By an extension of the method to include other factors the analysis may be further pursued.

From analyses made through the elimination of factors by partial correlation, "causal" relationships may often be determined. Phillips,* for example, in a study of causes contributing to absence on account of illness among government employees over a period of a year, found that the observed correlation between absence and mean temperature on the day

* Phillips, F. E., *Application of Partial Correlation to a Health Problem*, Public Health Report, Reprint No. 867, 1923.

of absence was $-.37$. When the four factors (1) relative humidity at 8:00 A.M.; (2) relative humidity at noon of the previous day; (3) inches of rainfall on the day of absence; and (4) percent of possible sunshine on the day of absence were held constant, the net correlation remaining between absence and temperature was $-.39$. Since the correlation between absence and temperature was the only r of any size (the other r 's, both entire and partial, were negligible) the conclusion seems to be that of the factors studied, temperature on the day of absence is the most important contributing cause of absence. Illness, of course, must be taken as the primary cause of absence. It must be clearly understood that partial correlation has nothing to say about causes as such. One cannot say which of two variables is the cause and which the effect, when all one has is the correlation between them. Sometimes, however, cause and effect distinctions are a matter of common-sense analysis. In the illustration given above, for instance, the distinction between cause and effect is obvious.

Another interesting example of the use of partial correlation in a causal investigation is supplied by the work of Reavis.* This investigator undertook to ferret out the causes of attendance and non-attendance in rural schools. Certain factors (1) distance from school, (2) age-grade relationship, (3) kind of work done by pupils, (4) training and experience of the teacher, (5) school equipment, and (6) character of the community were selected as having presumably some effect upon school attendance. When partial correlation coefficients were calculated it was found that the original correlations between attendance and distance from school, and between attendance and character of the community, were the least reduced. The first coefficient was lowered from $-.45$ to $-.43$; and the second from $.30$ to $.28$. Of all of the factors selected, these two seemed to have the most direct or independent influence

* Reavis, George, *Factors Controlling Attendance in Rural Schools*, Teachers College, Columbia University, Contributions to Education, 1920, No. 108, pp. 52-69.

upon school attendance. As in the problem of temperature and illness, cited above, the distinction here between cause and effect is clear — it is evident that distance from school and character of community are the causes and not the effects of good or poor school attendance.

2. The Value of Multiple Correlation in Analysis

Multiple correlation is especially useful when one wishes to determine the influence of a group of factors (e.g., battery of tests) upon some final result. Moreover, the calculation of successive R 's as tests are added to the team (p. 432) enables one to estimate the relative contribution of each factor. Only a few illustrations of the application of multiple correlation to psychological problems can be cited here; but the student will encounter many in the references given, and in the literature. Gates* has employed multiple correlation in an effort to encompass as many as possible of the factors which influence school achievement. In a group of 57 fourth-grade children, the r between educational achievement and M.A. was .595. When physical efficiency (vigor, stamina, etc.) as estimated by teachers was added to M.A., the R of educational achievement with M.A. plus physical efficiency was .653 — a gain of about .06 point. However, when emotional maturity (as estimated by teachers) was added to the battery M.A. plus physical efficiency, and still further social maturity (as estimated by teachers) was added to M.A. plus physical efficiency plus emotional maturity, the multiple correlation was unchanged. Gates concludes: "Physical fitness, then, appears to exert a greater specific influence (i.e., over and above the r with M.A.) upon achievement than does either social or emotional maturity or both combined. Both combined add practically nothing of value to a team of M.A. plus physical fitness for purposes of predicting scholastic achievement."

* Gates, A. I., *The Nature and Educational Significance of Physical Status and of Mental, Physiological, Social and Emotional Maturity*, Journal of Educational Psychology, 1924, 15, pp. 347-349.

In a study of the factors influencing children's answers to a series of health tests designed to measure knowledge of hygienic procedures, habits, attitudes, etc., Franzen* has made an interesting use of multiple correlation. Intelligence as measured by the McCall Multi-Mental Scale gave the highest correlation with the five health-knowledge tests. The addition of such factors as economic rating of home, cleanliness rating of child, grade, and age, added in order to the general intelligence test, did not appreciably change the correlation of the battery with the health test, the average increase being less than .05. The addition of a sixth variable called "nativity stock" did, however, raise the correlation. This last factor is described as an index of social status and was based upon the national origin of parents; it was determined from the intelligence test ratings of various national groups on Army Alpha obtained during the World War. On the basis of his results, Franzen concludes that apart from general intelligence the racial or national group to which the child belongs is the "strongest determinant of scores on the test."

Burks† has made use of multiple correlation in determining the relative contribution of heredity and environment to a child's I.Q. as measured by Stanford-Binet. The R between I.Q. and parental intelligence test score plus environmental index (by Whittier Home Scale) was found to be .61 for an N of 105. Since R^2 is .37, about 37% of the variance of children's intelligence (p. 432) may be attributed to the combined effect of home environment and parents' mental level. Parental intelligence contributed 33%, and home environment 4%, to the 37% accounted for by these two factors. The remaining 63% is attributable to factors not measured by these two.

* Franzen, R., *Health Education Tests*, School Health Research Monograph, American Child Health Association, 1929, No. 1.

† Burks, B. S., *The Relative Influence of Nature and Nurture upon Mental Development, a Comparative Study of Foster Parent-Foster Child Resemblance and True Parent-True Child Resemblance*, 27th Yearbook, N.S.S.E., 1928, Part I, pp. 219-316.

3. The Value of the Regression Equation in Prediction and Analysis

The regression equation may be written in two ways. In its usual form (see formula (116), p. 426) it gives the weights to be assigned to the scores of each of the several independent variables, $X_2, X_3 \dots X_n$, in order that X_1 may be predicted or forecast, by the given battery, with minimum error. When the scores in the independent variables are expressed as standard or z -scores, the regression equation enables us to predict our criterion in standard score form. What is more important, however, the standard score form of the regression equation permits us to analyze, within certain limits, the capacity or ability measured by X_1 in terms of the contributions of the independent variables. In the present section we shall illustrate in order the use of the regression equation (1) in usual score form, and (2) in standard score form.

(1) We have already learned (p. 430) that the regression equation enables one to combine two or more tests or other measures (independent variables $X_2, X_3, \dots X_n$) into a single value which will give the best possible estimate of X_1 (the dependent variable). In the three-variable problem on page 415, for example, the regression equation gives us the best possible forecast of the number of honor points which a student will receive when we know his general intelligence score and the average number of hours he spends in study per week. Once calculated, the regression equation may be used subsequently to estimate the scores of students not included in the original group, if their scores in the general intelligence test and the average number of hours spent in study per week are known. The accuracy of the regression equation as a forecasting instrument is determined (p. 443) by the size of the standard error of estimate and by the multiple coefficient of correlation.

A good illustration of the value of the regression equation in forecasting, taken from the field of agriculture, may be found

in the work of Moore* in forecasting the cotton crop in the Southern States. Taking the cotton crop in Georgia as the dependent variable (to cite a single example) and the May rainfall, June temperature and August temperature as independent variables, Moore built up a regression equation from which it was possible to get a better forecast of the cotton crop at the end of August than the official method of the U.S. Department of Agriculture could obtain from the condition of the crop in September.

In addition to its use as a forecasting instrument, the regression equation may be used also to determine the weight which each test in a battery should have in order that the composite score obtained from the battery shall be the best possible estimate of the criterion which the whole battery presumably measures. Except for the fact that the primary interest here lies in the weights which the independent variables take in the regression equation, this is essentially the same problem as that of prediction or forecasting discussed in the last few paragraphs. To illustrate, suppose that we wish to devise a group test for measuring general intelligence; and that we wish this battery to consist of four tests. The first step is to obtain as many valid criteria of general intelligence as possible. These may be school grades, teachers' estimates, other standard intelligence examinations, as for example Stanford-Binet and Army Alpha, and various combinations of these. The next step is to select four tests which will have (1) high correlations with the criteria and (2) low correlations with each other.† If these two requirements are satisfied or approximately satisfied, each test will not only measure some aspect of the criteria, but will measure a different or slightly different phase, since the low intercorrelations prevent much duplication. Let us call the criterion X_c and the four tests of the battery X_1 , X_2 , X_3 and X_4 .

* Moore, H. L., *Forecasting the Yield and Price of Cotton*, 1917, pp. 108-115.

† The ideal battery consists of tests which correlate as highly as possible with the criteria and as low as possible with each other (see p. 445).

The regression equation in score form will then become $\bar{X}_c = AX_1 + BX_2 + CX_3 + DX_4 + K$. In this regression equation A , B , C , and D are the regression coefficients or weights to be given the scores made on the four tests, while K is a numerical constant (p. 426). To take a hypothetical case, suppose that $A = 1$, $B = 2$, $C = 3$, and $D = 4$. The regression equation would then become $\bar{X}_c = 1X_1 + 2X_2 + 3X_3 + 4X_4 + K$; and a subject's score on Test 1 must be multiplied by 1, his score on Test 2 by 2, his score on Test 3 by 3, and his score on Test 4 by 4 in order that the composite score on the battery may give the best estimate of his score in X_c , the criterion.

The regression equation furnishes the "ideal" method (theoretically) of combining several tests into a team, as each test is weighted with respect to its correlation with the criterion, independently of the other tests in the team. Under these conditions the standard error of estimate is a minimum, while the correlation (multiple R) of the predicted X_c scores and the actual X_c scores is the maximum obtainable with the given set of tests.

(2) The difference between the ordinary form of the regression equation and the standard score form now to be considered is that in the latter all variables are expressed as standard scores (p. 178). The σ 's of all the tests, therefore, are 1, and the means of all the tests are .00. The use of standard scores eliminates differences in the size of test units as well as differences in variability. Hence we are able to determine *from the correlation alone* the relative weight with which each independent factor "enters into" or contributes to the dependent variable, independently of the other factors. This permits an analysis of the relative importance of the different independent factors in accounting for a final result.

We may illustrate the regression equation expressed in standard scores using the data taken from the three-variable problem on page 415. If X_1 , honor points, is taken as the "cri-

terion," while X_2 , general intelligence, and X_3 , average number of hours spent in study per week are, as before, the independent variables, the regression equation [formulas (115) and (117)] is written

$$\bar{x}_1 = r_{12.3} \frac{\sigma_{1.23}}{\sigma_{2.13}} x_2 + r_{13.2} \frac{\sigma_{1.23}}{\sigma_{3.12}} x_3 \quad (\text{deviation form})$$

Expanding the partial σ 's by formula (114)

$$\begin{aligned} \bar{x}_1 = r_{12.3} \frac{\sigma_1 \sqrt{1-r_{13}^2} \sqrt{1-r_{12.3}^2}}{\sigma_2 \sqrt{1-r_{23}^2} \sqrt{1-r_{12.3}^2}} x_2 \\ + r_{13.2} \frac{\sigma_1 \sqrt{1-r_{12}^2} \sqrt{1-r_{13.2}^2}}{\sigma_3 \sqrt{1-r_{23}^2} \sqrt{1-r_{13.2}^2}} x_3 \end{aligned}$$

If we divide through by σ_1 and write σ_2 and σ_3 under x_2 and x_3 , respectively, this equation, upon cancelling the radical expressions, becomes

$$\frac{\bar{x}_1}{\sigma_1} = r_{12.3} \frac{\sqrt{1-r_{13}^2}}{\sqrt{1-r_{23}^2}} \frac{x_2}{\sigma_2} + r_{13.2} \frac{\sqrt{1-r_{12}^2}}{\sqrt{1-r_{23}^2}} \frac{x_3}{\sigma_3}$$

Now putting z_1 equal to x_1/σ_1 , z_2 equal to x_2/σ_2 and z_3 equal to x_3/σ_3 we have

$$\bar{z}_1 = r_{12.3} \frac{\sqrt{1-r_{13}^2}}{\sqrt{1-r_{23}^2}} z_2 + r_{13.2} \frac{\sqrt{1-r_{12}^2}}{\sqrt{1-r_{23}^2}} z_3 \quad (125)$$

(regression equation for three variables in terms of standard scores)

If we substitute in formula (125) the numerical values found in Table 60 we have

$$\begin{aligned} \bar{z}_1 &= .80 \frac{.9474}{.9367} z_2 + .71 \frac{.8000}{.9367} z_3 \\ &= .81z_2 + .60z_3 \end{aligned}$$

This equation may be interpreted to mean that insofar as the two factors, general intelligence and number of hours spent in study per week, *contribute* to the ability to earn honor points, their relative weights are .81 : .60 or 4 : 3. This ratio gives the relative contributions of the two measured abilities independent of the test units, and not the relative weights of their *scores*.

The weight to be assigned each score is found from the usual regression equation given on page 426. It is of interest to note that while the scores on the general intelligence test and number of study hours per week are as 1:2 (p. 418), the actual contribution of these measures to honor points is as 4:3. As we should expect, intelligence as measured by the given tests has more weight than hours spent in study in determining the number of honor points earned.

The general regression equation for n variables, written in standard scores, is

$$\bar{z}_1 = \beta_{12.34 \dots n} z_2 + \beta_{13.24 \dots n} z_3 + \dots + \beta_{1n.23 \dots (n-1)} z_n \quad (126)$$

*(regression equation, deviation form, for n variables
in terms of standard scores)*

The symbol β is used to distinguish the partial regression coefficients in terms of standard scores from the partial regression coefficients (b 's) in terms of the scores of the test. The β 's may be obtained from the b 's by means of the following formula:

$$\beta_{12.34 \dots n} = b_{12.34 \dots n} \frac{\sigma_2}{\sigma_1} \quad (127)$$

*(standard score partial regression coefficients (β 's) expressed
in terms of test score regression coefficients (b 's))*

Regression coefficients in terms of β are usually called "beta weights," to distinguish them from the "score weights" (the b 's) in the ordinary regression equation. In the three-variable problem above, the beta weight of general intelligence (2) is .81 and the beta weight of study hours (3) is .60. The beta weights, as already indicated, give the contributions of the various independent variables in the regression equation to the dependent variable (the criterion). They are very useful in enabling us to analyze the comparative worth of several factors insofar as they bear upon some given result.

4. Limitations to the Use of Partial and Multiple Correlation

In concluding this discussion of partial and multiple correlation, certain limitations to the use of the method should be pointed out (1) In the first place, in order that partial coefficients of correlation be valid measures of relationship, it is necessary that all zero order coefficients be computed from data in which the regression is linear. If there is any doubt as to linearity, the tests given on page 400 should be employed. (2) Secondly, the number of cases in a multiple correlation problem should be large, especially if there are a number of variables; otherwise the coefficients calculated from the data will have little significance. Coefficients which are misleadingly high or low may be obtained when studies which involve many variables are based on relatively few cases.* The question of accuracy of computation is also involved here. A general rule advocated by many workers is that results should be carried to as many decimals as there are variables in the problem. Whether or not this rule is to be followed, however, must depend upon the accuracy of the original measures (3) A serious limitation to a clear-cut interpretation of a partial r arises from the fact that most tests employed by psychologists probably depend upon a large number of "psychological determiners." When we "partial out" the influence of such clear-cut and relatively objective factors as age, height, school grade, etc., we have a reasonably clear notion of what the partial r 's mean. But when we attempt to render "logical memory" constant, say, by partialling out memory test scores from the correlation between general intelligence and educational achievement, the result is by no means so unequivocal. The abilities determining the scores in general intelligence *and* in school achievement undoubtedly overlap the memory test in other respects than in the "memory" involved. Partialling out a memory test score from the correlation between general intelligence and educational achievement, therefore, will render constant the influence of

* Paterson, D. G., *Physique and Intellect*, 1930, p. 47.

many factors not strictly "memory," i.e., partial out too much.*

To illustrate again the difficulty which arises when partial correlation is applied to test batteries: it would be fallacious to interpret the partial correlation between reading comprehension and arithmetic, say, with the influence of "general intelligence" partialled out, as giving the net relationship between these two variables for a "constant" degree of intelligence. Both reading and arithmetic enter with heavy, but unknown, weight into most general intelligence tests; hence the partial correlation between these two, for general intelligence constant, cannot be interpreted in a clear-cut and meaningful way.

Partial r 's obtained from psychological and educational tests, though often difficult to interpret, may be used in multiple regression equations when the purpose is to determine the relative weight to be assigned the various tests of a battery. But we should be cautious in attempting to give psychological meaning to such residual, i.e., partial, r 's. Several writers have discussed this problem, and should be referred to by the investigator who plans to use partial and multiple correlation extensively.†

(4) Perhaps the chief limitation to R , the coefficient of multiple correlation, is the fact that, since it is always positive, variable errors of sampling tend to accumulate and thus make the coefficient too large. A formula for computing a "chance R ," with which the obtained R may be compared, was given on page 431. A correction to be applied to R , when the sample is small and the number of variables large, has been given by Ezekiel‡ This correction adjusts the R for "shrinkage"—gives the value which R would most probably take in the

* Burks, B. S., *Statistical Hazards in Nature-Nurture Investigations*, 27th Yearbook, N.S.S.E., 1928, Part I, pp. 9-33.

† Burks, B. S., *On the Inadequacy of the Partial and Multiple Correlation Technique*, *Journal of Educational Psychology*, 1926, 17, pp. 532-540.

Hull, C. L., *Aptitude Testing*, 1928, pp. 250-253.

Moore, T. V., *Partial Correlations*, *Studies in Psychology and Psychiatry from the Catholic University of America*, 1932, 3, pp. 1-39.

‡ Ezekiel, M., *Methods of Correlation Analysis*, 1930, pp. 176-177.

population from which our sample was drawn. Unless the sample is small (25 or less) and the number of variables large, however, this correction is negligible, and may be safely disregarded.

VII. SPURIOUS CORRELATION

The correlation between two sets of test scores is said to be *spurious* when it is due in some part, at least, to factors other than those which determine performance in the tests themselves. In general, the cause of spurious correlation may be said to lie in a failure to control conditions; and the most usual effect of this lack of control is a "boosting" or dilation of the coefficient. Some of the situations which may lead to spurious correlation will be given in this section.

1. Spurious Correlation Arising from Heterogeneity

We have shown elsewhere how a lack of uniformity in age conditions will lead to correlations which are spuriously high. Failure to take account of heterogeneity introduced by the age factor is a prolific source of error in correlational work. To cite an example, within a group of boys 10 to 18 years old, a substantial correlation will appear between strength of grip and memory span, quite apart from any intrinsic relation, due solely to the fact that both attributes tend to increase with age. In stating the correlation between two tests, or the reliability coefficient of a test (p. 322) one should always be careful to specify the range of ages, grades included, and other data bearing upon physical, mental, and cultural differences, in order to show the degree of heterogeneity in the group. Without this information, the r itself is of little value.

Heterogeneity is introduced by other factors than age. If alcoholism, degeneracy, and bad heredity are all positively related, the r between alcoholism and degeneracy will be too high (because of the effect of heredity upon both factors) unless heredity can be "held constant." Again, suppose that we have measured two distinctly different groups, 500 college seniors,

and 500 day laborers, upon a cancellation test and upon a general intelligence test. The mean ability in both tests will be definitely higher in the college group. Now even if the correlation between the two tests is zero within each group taken separately, if the two groups are combined, a positive correlation will appear which is attributable to the heterogeneity of the group with respect to age, intelligence, and educational background. Such a correlation is, of course, spurious.*

To be a valid measure of relationship, a correlation coefficient must be freed of the extraneous influences which affect the relationship between the variables concerned. This may be accomplished (1) by selecting samples or groups in which age (or whatever the factor to be controlled) is constant. This is the method of experimental control (p. 410). (2) One may use partial correlation if the factor to be controlled can be measured and its correlation with the variables studied can be found.

2. Spurious Index Correlation †

Even when three variables X_1 , X_2 , and X_3 are uncorrelated, a correlation between the indices Z_1 and Z_2 (where $Z_1 = X_1/X_3$, and $Z_2 = X_2/X_3$) may appear which is as large as .50. To illustrate, if two individuals observe a series of magnitudes (e.g., Galton bar settings) independently, the absolute errors of observation (X_1 and X_2) may be uncorrelated, and still an appreciable correlation appear between the *errors* made by the two observers, when these are expressed as *percents* of the observed magnitude (X_3). The spurious element here, of course, is the common factor X_3 in the denominator of the ratios.

One of the commonest examples of spurious index correlation in psychology is found in the correlation of I.Q.'s obtained from two different tests. If the I.Q.'s of 500 children ranging in age

* Garrett, H. E., and Anastasi, A., *The Tetrad-Difference Criterion and the Measurement of Mental Traits*, Annals New York Academy of Sciences, 1932, 33, pp. 233-282

† Yule, G. U., *An Introduction to the Theory of Statistics*, 1929, pp. 215-216.

Thomson, G. H., and Pintner, R., *Spurious Correlation and Relationship Between Tests*, Journal of Educational Psychology, 1924, 15, pp. 433-444.

from 3 to 14 years are calculated from two tests X_1 and X_2 , the correlation between the I.Q.'s $\left(\text{I.Q.} = \frac{\text{M.A.}}{\text{C.A.}} \right)$ obtained from the two tests will be spurious owing to the common factor of chronological age in both series.

3. Spurious Correlation between Averages

Spurious correlation usually results when the average scores made by a number of different groups on a given test are correlated against the average scores made by the same groups on a second test. An example is furnished by the correlations reported by Bagley* between the mean Army Alpha scores, by states, and such "educational" factors as number of schools, books sold, magazines circulated in the states, etc. Most of these correlations are high—many above .90. If such average correlations are compared with the correlations between the intelligence scores made by *individuals* and number of years spent in school, these latter are usually lower—around .60. Correlations between averages become "inflated" because a large number of factors which would ordinarily reduce the correlation between intelligence and education cancel out when averages are taken. Bagley's high correlations do not mean, therefore, that the relationship between intelligence and education is as high as his average correlations represent it to be.

4. Spurious Correlation Resulting from the Correlation of a Single Test with a Composite of which it is a Member

If the scores of several tests, X_1, X_2, X_3 , etc., are averaged or added, and the composite scores, $X_{\text{com.}}$, are correlated against the scores of any single test, the resulting relationship will be spurious because of the presence of X_1 in the composite. The amount or degree of the spurious element is measured by the ratio $\sqrt{t/s}$ in which t = the number of elements in the single test, and s = the number of elements in the composite.† To

* Bagley, W. C., *Determinism in Education*, 1925, p. 81.

† Musselman, J. R., *Spurious Correlation Applied to Urn Schemata*, Journal of the American Statistical Association, 1923, 18, pp. 908-911.

illustrate, there are 20 elements in the number series completion tests of the Army Alpha and 212 items in the whole test. If there were no correlation between the scores in the number series completion test and the Alpha there would still be a correlation between the two sets of scores equal to the square root of the ratio found by dividing the number series completion items by the items in Alpha—i.e., $\sqrt{20/212}$ or .30. A correlation obtained between number series completion and Alpha, therefore, will be too high because of the inclusion of the number series completion items in both sets of scores. (p. 350)

When several tests are all of the same, or approximately the same length, the amount of spurious relation which will result from the correlation of each *single* test with the *composite* of them all will be approximately constant ($\sqrt{t/s}$ is constant). Hence, it would be valid to compare the correlations of the separate tests with the composite if we wish to discover which tests are most representative of the ability measured in common by them all. Since all r 's are equally spurious, they are directly comparable.

PROBLEMS

1. The correlation between a general intelligence test and school achievement in a group of children from eight to fourteen years old is .80. The correlation between the general intelligence test and age in the same group is .70; and the correlation between school achievement and age is .60. What is the correlation between general intelligence and school achievement in children of the same age? Comment upon your result.
2. In a group of 100 college freshmen, the correlation between (1) Army Alpha and (2) the A-cancellation test is .20. The correlation between (1) Army Alpha and (3) a battery of controlled association tests in the same group is .70. If the correlation between (2) cancellation and (3) controlled association is .45, what is the "net" correlation between Army Alpha and cancellation in this group? Between Alpha and controlled association? Interpret your results.

3. Explain why some variables are of such a nature that it is difficult to hold them "constant," and hence to employ them in problems involving partial correlation.
4. Given the following data.

X_1 = Stanford-Binet I Q.		
X_2 = Memory for Objects		
X_3 = Cube Imitation		
$M_1 = 101.71$	$M_2 = 10.06$	$M_3 = 3.35$
$\sigma_1 = 13.65$	$\sigma_2 = 3.06$	$\sigma_3 = 2.02$
$r_{12} = .41$	$r_{13} = .50$	$r_{23} = .16$

- (a) Work out the regression equation of X_1 upon X_2 and X_3 .
- (b) Compute $R_{1(23)}$ and $\sigma_{(\text{est } X_1)}$.
- (c) If a child's score is 12 in Test X_2 and 4 in Test X_3 , what is his most probable score in X_1 (I.Q.) and the $PE_{(\text{est } X_1)}$?
5. Given the following data* :

(1)	(2)	(3)	(4)	(5)
Comprehension of Literature (Stanford)	Reading Speed	Grammar	Information Range	Vocabulary
$M_1 = 63$	$M_2 = 96$	$M_3 = 35$	$M_4 = 96$	$M_5 = 92$
$\sigma_1 = 22$	$\sigma_2 = 30$	$\sigma_3 = 6$	$\sigma_4 = 20$	$\sigma_5 = 30$
$r_{12} = .33$	$r_{23} = .37$	$r_{34} = .64$	$r_{45} = .74$	
$r_{13} = .42$	$r_{24} = .40$	$r_{35} = .58$		
$r_{14} = .65$	$r_{25} = .57$			
$r_{15} = .69$				

(In the following round off all results including $\sqrt{1 - r^2}$ to two decimals.)

- (a) With (1) Comprehension as the "criterion" or dependent variable, work out a regression equation involving variables 1, 2, and 3. Compute $R_{1(23)}$ and $\sigma_{(\text{est } 1)}$.
- (b) With (1) Comprehension as the criterion, work out a regression equation involving the first four variables. Compute $R_{1(234)}$ and $\sigma_{(\text{est } 1)}$.

* Adapted from Peters, Charles C., and VanVoorhis, Walter R., *Statistical Procedures and Their Mathematical Bases*, 1935, pp. 196-200.

- (c) With (1) still as the criterion, work out a regression equation involving all five variables. Compute $R_{1(2345)}$, and $\sigma_{(\text{est } 1)}$
- (d) What are the relative weights of tests 2, 3, 4, and 5 in determining Comprehension of Literature? (Hint: Express the regression equation obtained in (c) in standard scores)
- (e) Find what percentage of the variance of variable (1) is determined by tests 2 and 3 together; by tests 2, 3, and 4 together; and by tests 2, 3, 4, and 5 taken together (p 432).
6. Suppose that a student makes the following scores on four of the tests described in Problem 5.

Reading = 100	Information = 90
Grammar = 30	Vocabulary = 80

What is this student's most probable score in Comprehension?

7. Let X_1 be a criterion and X_2 and X_3 be two other tests. Correlations and σ 's are as follows:

$r_{12} = .60$	$\sigma_1 = 5.00$
$r_{13} = .50$	$\sigma_2 = 10.00$
$r_{23} = .20$	$\sigma_3 = 8.00$

- (a) How much more accurately can X_1 be predicted from X_2 and X_3 than from either alone?
- (b) If N is 100, compare the "chance" R with the obtained $R_{1(23)}$ (p. 431).
8. Given a team of two tests, each of which correlates .50 with a criterion. If the two tests correlate .20
- (a) How much would the addition of another test which correlates .50 with the criterion and .20 with each of the other tests improve the predictive value of the team?
- (b) How much would the addition of two such tests improve the predictive value of the team?
9. Two absolutely independent tests B and C completely determine the criterion A . If B correlates .50 with A , what is the correlation of C and A ? What is the multiple correlation of A with B and C ?
10. A group test contains ten sub-tests. One of these sub-tests correlates .60 with total scores on the test. If there are 200 items in the test, and 15 items in the sub-test, what correlation might be expected between the test and the sub-test in the absence of any actual correlation?

11. Comment upon the following statements:

- (a) It is good practice to correlate I.Q.'s achieved upon two general intelligence tests, no matter how wide the age range.
- (b) The positive correlation between average Army Alpha scores by states and the number of libraries in the states proves the close dependence of Alpha upon education.
- (c) The correlation between intelligence test scores and tapping rate in a group of 200 eight year old children is .20; and the correlation between intelligence test scores and tapping rate in a group of 100 college freshmen is .10. However, when the two groups are combined the correlation between these two tests becomes .40. This shows that we must have a large group in order to get high correlations.

ANSWERS

1. $r = .67$
2. $r(\text{Alpha and cancellation}) = -.19$; $r(\text{Alpha and controlled association}) = .70$
4. (a) $\bar{X}_1 = 1.47X_2 + 2.98X_3 + 76.94$
 (b) $R_{1(23)} = .60$, $\sigma_{(\text{est } X_1)} = 10.93$
 (c) 106.50 ± 7.37
5. (a) $\bar{X}_1 = .14X_2 + 1.26X_3 + 5$
 $R_{1(23)} = .48$, $\sigma_{(\text{est } X_1)} = 19$
 (b) $\bar{X}_1 = .06X_2 + .00X_3 + .68X_4 - 8$
 $R_{1(234)} = .66$; $\sigma_{(\text{est } X_1)} = 17$
 (c) $\bar{X}_1 = -.05X_2 - .30X_3 + .38X_4 + .38X_5 + 7$
 $R_{1(2345)} = .73$; $\sigma_{(\text{est } X_1)} = 15$
 (d) $\bar{z}_1 = -.07z_2 - .08z_3 + .35z_4 + .52z_5$
 (e) 23%; 44%; 53%
6. 57
7. (a) From X_2 alone, $\sigma_{(\text{est } X_1)} = 4.0$
 From X_3 alone, $\sigma_{(\text{est } X_1)} = 4.3$
 From X_2 and X_3 , $\sigma_{(\text{est } X_1)} = 3.5$
 (b) $R_{1(23)} = .71$. The "chance" $R = .14$
8. (a) R increases from .64 to .73
 (b) R increases from .64 to .79
9. $r_{AC} = .87$; $R_{A(BC)} = 1.00$
10. .27

REFERENCES

The following books will be found useful by students of psychology and education:

Statistical Methods Applied to Education, by Harold O. Rugg. Houghton Mifflin Co. 1917.

The Theory of Educational Measurements, by Walter Scott Monroe. Houghton Mifflin Co. 1923.

The Fundamentals of Statistics, by L. L. Thurstone. The Macmillan Co. 1925.

Statistical Method in Educational Measurement, by Arthur S. Otis. World Book Co. 1925.

The Interpretation of Educational Measurements, by T. L. Kelley. World Book Co. 1927.

Statistical Methods for Students in Education, by Karl J. Holzinger. Ginn and Co. 1928.

Mathematics Essential for Elementary Statistics, by Helen M. Walker. Henry Holt and Co. 1934.

Statistical Method in Education, by Charles W. Odell. Appleton-Century Co. 1935.

Statistics for Students of Psychology and Education, by Herbert Sorenson. McGraw-Hill Co. 1936.

Psychometric Methods, by J. P. Guilford. McGraw-Hill Co. 1936.

More advanced books are:

Essentials of Mental Measurement, by William Brown and Godfrey H. Thomson. Cambridge University Press. 1921.

A First Course in Statistics, by D. Caradog Jones. G. Bell & Sons. 1921.

Statistical Method, by Truman L. Kelley. The Macmillan Co. 1923.

Handbook of Mathematical Statistics, by H. L. Rietz and others. Houghton Mifflin Co. 1924.

Medical Biometry and Statistics, by Raymond Pearl. Saunders. 1930 (second edition).

Statistical Methods for Research Workers, by R. A. Fisher. Oliver and Boyd. 1936 (sixth edition).

Methods of Correlation Analysis, by Mordecai Ezekiel. John Wiley & Sons. 1930.

An Introduction to the Theory of Statistics, by G. Udny Yule. Chas. Griffin and Co. 1932 (10th edition).

Statistical Procedures and their Mathematical Bases, by Charles C. Peters and Walter R. VanVoorhis. Pennsylvania State College. 1935.

Statistical Methods in Biology, Medicine, and Psychology, by C. B. Davenport and Merle P. Ekas. John Wiley and Sons. 1936.

Computation Aids:

Barlow's Tables of Squares, Cubes, Square Roots, Reciprocals of Numbers from 1 to 10,000. E. and F. N. Spon, Ltd. 1921.

Tables of $\sqrt{1-r^2}$ and $1-r^2$ for Use in Partial Correlation and Trigonometry, by John Rice Miner. Johns Hopkins Press. 1922.

Standard Table of Square Roots, by L. M. Milne-Thomson. G. Bell & Sons. 1929.

Statistical Tables for Students in Education and Psychology, by Karl J. Holzinger. University of Chicago Press. 1925

Handbook of Statistical Nomographs, Tables, and Formulas, by Jack W. Dunlap and Albert K. Kurtz. World Book Co. 1932.

Study Manuals:

Work Book in Educational Measurements, by Harry A. Greene. Longmans. 1928.

Study Manual in Elementary Statistics, by Everett F. Lindquist and George D. Stoddard. Longmans. 1929.

TABLE 16

FRACTIONAL PARTS OF THE TOTAL AREA (TAKEN AS 10,000) UNDER THE NORMAL PROBABILITY CURVE, CORRESPONDING TO DISTANCES ON THE BASELINE BETWEEN THE MEAN AND SUCCESSIVE POINTS LAID OFF FROM THE MEAN IN UNITS OF PE

Example: we find between the mean and a point $1.55 PE$ ($\frac{x}{PE} = 1.55$) from the mean 35.21% of the entire area under the curve.

$\frac{x}{PE}$.00	.05	$\frac{x}{PE}$.00	.05
.0	0000	0135	3.0	4785	4802
.1	0269	0403	3.1	4817	4832
.2	0537	0670	3.2	4846	4858
.3	0802	0933	3.3	4870	4881
.4	1063	1193	3.4	4891	4900
.5	1320	1447	3.5	4909	4917
.6	1571	1695	3.6	4924	4931
.7	1816	1935	3.7	4937	4943
.8	2053	2168	3.8	4948	4953
.9	2281	2392	3.9	4957	4961
1.0	2500	2606	4.0	4965	4968
1.1	2709	2810	4.1	4972	4974
1.2	2909	3004	4.2	4977	4979
1.3	3097	3187	4.3	4981	4983
1.4	3275	3360	4.4	4985	4987
1.5	3442	3521	4.5	4988	4989
1.6	3597	3671	4.6	4990	4991
1.7	3742	3811	4.7	4992	4993
1.8	3876	3939	4.8	4994	4995
1.9	4000	4058	4.9	4995	4996
2.0	4113	4166	5.0	4996	4997
2.1	4217	4265	5.1	4997.1	4997.4
2.2	4311	4354	5.2	4997.7	4998
2.3	4396	4435	5.3	4998.2	4998.5
2.4	4473	4508	5.4	4998.6	4998.8
2.5	4541	4573	5.5	4999	4999.1
2.6	4603	4631	5.6	4999.2	4999.3
2.7	4657	4682	5.7	4999.4	4999.5
2.8	4705	4727	5.8	4999.54	4999.6
2.9	4748	4767	5.9	4999.65	4999.7

TABLE 34

TO FIND THE CHANCES OF A SIGNIFICANT DIFFERENCE, I.E., TO FIND THE CHANCES THAT THE TRUE DIFFERENCE IS GREATER THAN ZERO, GIVEN THE OBTAINED DIFFERENCE BETWEEN TWO MEASURES, AND THE STANDARD ERROR OF THE DIFFERENCE

Example: A D/σ_D of 1.3 means that the chances are 90 in 100 that the obtained difference is significant — that the true difference is greater than zero.

$\frac{D}{\sigma_D}$	Chances in 100	$\frac{D}{\sigma_D}$	Chances in 100
.00	50	1.15	87
.05	52	1.20	88
.10	54	1.25	89
.15	56	1.30	90
.20	58	1.35	91
.25	60	1.40	92
.30	62	1.45	93
.35	64	1.50	93
.40	65	1.60	94
.45	67	1.70	96
.50	69	1.80	96
.55	71	1.90	97
.60	73	2.00	98
.65	74	2.10	98
.70	76	2.20	99(98.6)
.75	77	2.30	99(98.9)
.80	79	2.40	99(99.2)
.85	80	2.50	99(99.4)
.90	82	2.60	99(99.5)
.95	83	2.70	100(99.7)
1.00	84	2.80	100(99.74)
1.05	85	2.90	100(99.8)
1.10	86	3.00	100(99.9)

TABLE 35

TO FIND THE CHANCES OF A SIGNIFICANT DIFFERENCE, GIVEN
THE OBTAINED DIFFERENCE BETWEEN TWO MEASURES AND
THE PROBABLE ERROR OF THAT DIFFERENCE

Example: A D/PE_D of 1.10 means that there are 77 chances in 100 that the obtained difference is significant, namely, that the true difference is greater than zero.

$\frac{D}{PE_D}$	Chances in 100	$\frac{D}{PE_D}$	Chances in 100
.00	50	1.55	85
.05	51	1.60	86
.10	53	1.65	87
.15	54	1.70	87
.20	55	1.75	88
.25	57	1.80	89
.30	58	1.85	89
.35	59	1.90	90
.40	61	1.95	91
.45	62	2.00	91
.50	63	2.10	92
.55	64	2.20	93
.60	66	2.30	94
.65	67	2.40	95
.70	68	2.50	95
.75	69	2.60	96
.80	71	2.70	97(96.6)
.85	72	2.80	97
.90	73	2.90	97(97.5)
.95	74	3.00	98(97.9)
1.00	75	3.10	98
1.05	76	3.20	98(98.5)
1.10	77	3.30	99(98.7)
1.15	78	3.40	99(98.9)
1.20	79	3.50	99
1.25	80	3.60	99
1.30	81	3.70	99
1.35	82	3.80	99(99.5)
1.40	83	3.90	100(99.6)
1.45	84	4.00	100(99.7)
1.50	84		

TABLE 41

PROBABLE ERRORS OF THE COEFFICIENT OF CORRELATION FOR VARIOUS
NUMBERS OF MEASURES (N) AND FOR VARIOUS VALUES OF r

Number of Measures	Correlation Coefficient r						
	0 00	0 10	0 20	0 30	0 40	0 50	0 60
20	1508	1493	1448	1373	1267	1131	0965
30	1231	1219	1182	1121	1035	0924	0788
40	1067	1056	1024	0971	0896	0800	0683
50	0954	0944	0915	0868	0801	0715	0610
70	0806	0798	0774	0734	0677	0605	0516
100	0674	0668	0648	0614	0567	0506	0432
150	0551	0546	0529	0501	0463	0413	0352
200	0477	0472	0458	0434	0401	0358	0305
250	0426	0421	0409	0387	0358	0319	0272
300	0389	0386	0374	0354	0327	0292	0249
400	0337	0334	0324	0307	0283	0253	0216
500	0302	0299	0290	0274	0253	0226	0193
1000	0213	0211	0205	0194	0179	0160	0137
Number of Measures	0 65	0 70	0 75	0 80	0 85	0 90	0 95
20	0871	0769	0660	0543	0419	0287	0147
30	0711	0628	0539	0444	0342	0234	0120
40	0616	0544	0467	0384	0296	0203	0104
50	0551	0486	0417	0343	0265	0181	0093
70	0466	0411	0353	0290	0224	0153	0079
100	0391	0345	0294	0242	0187	0128	0066
150	0318	0281	0241	0198	0153	0105	0054
200	0275	0243	0209	0172	0133	0091	0047
250	0246	0218	0187	0154	0118	0081	0042
300	0225	0199	0170	0140	0108	0074	0038
400	0195	0172	0148	0122	0094	0064	0033
500	0174	0154	0132	0109	0084	0057	0029
1000	0123	0109	0093	0077	0059	0041	0021

TABLE 49

DEVIATES (x/σ) IN TERMS OF σ -UNITS AND ORDINATES (z) FOR
 GIVEN AREAS MEASURED FROM THE MEAN OF A NORMAL
 DISTRIBUTION WHOSE TOTAL AREA = 1.00
 $[x/\sigma = x]$

Area from the Mean (α)	x or (x/σ)	z	Area from the Mean (α)	x or (x/σ)	z
.00	.000	.399	.26	.706	.311
.01	.025	.399	.27	.739	.304
.02	.050	.398	.28	.772	.296
.03	.075	.398	.29	.806	.288
.04	.100	.397	.30	.842	.280
.05	.126	.396	.31	.878	.271
.06	.151	.394	.32	.915	.262
.07	.176	.393	.33	.954	.253
.08	.202	.391	.34	.995	.243
.09	.228	.389	.35	1.036	.233
.10	.253	.386	.36	1.080	.223
.11	.279	.384	.37	1.126	.212
.12	.305	.381	.38	1.175	.200
.13	.332	.378	.39	1.227	.188
.14	.358	.374	.40	1.282	.176
.15	.385	.370	.41	1.341	.162
.16	.412	.366	.42	1.405	.149
.17	.440	.362	.43	1.476	.134
.18	.468	.358	.44	1.555	.119
.19	.496	.353	.45	1.645	.103
.20	.524	.348	.46	1.751	.086
.21	.553	.342	.47	1.881	.068
.22	.583	.337	.48	2.054	.048
.23	.613	.331	.49	2.326	.027
.24	.643	.324	.50	∞	.000
.25	.675	.318			

TABLE 52. Giving the Probability (P) that, with a Given "Degree of Freedom," (n) the χ^2 Value Obtained in the Comparison of the Distribution of a Sample with that of a Theoretical Series Indicates that the Sample Belongs to or Has Arisen Out of Such Series. (The values of χ^2 are printed in the body of the table.) For larger values of n , the expression $\sqrt{2\chi^2} - \sqrt{2n} - 1$ may be used as a normal deviate with unit standard error

Adapted from R. A. Fisher's *Statistical Method for Research Workers*, Oliver & Boyd, by permission of publishers

n	$P = 0.99$	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0.000157	0.000628	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	0.115	0.185	0.362	0.584	1.005	1.424	2.368	3.665	4.642	6.251	7.815	9.837	11.341
4	0.297	0.429	0.711	1.064	1.649	2.195	3.351	4.787	5.989	7.779	9.488	11.668	13.277
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.503	10.645	12.592	15.033	16.812
7	1.239	1.564	2.187	2.833	3.822	4.671	6.346	8.383	9.808	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.584	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.568	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.265	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.862	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.787	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.125	16.114	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

TABLE 61

A TABLE TO INFER THE VALUE OF $\sqrt{1-r^2}$ FROM A
GIVEN VALUE OF r

r	$\sqrt{1-r^2}$	r	$\sqrt{1-r^2}$	r	$\sqrt{1-r^2}$
.0000	1.0000	.3400	.9404	.6800	.7332
.01	.9999	.35	.9367	.69	.7238
.02	.9998	.36	.9330	.70	.7141
.03	.9995	.37	.9290	.71	.7042
.04	.9992	.38	.9250	.72	.6940
.05	.9987	.39	.9208	.73	.6834
.06	.9982	.40	.9165	.74	.6726
.07	.9975	.41	.9121	.75	.6614
.08	.9968	.42	.9075	.76	.6499
.09	.9959	.43	.9028	.77	.6380
.10	.9950	.44	.8980	.78	.6258
.11	.9939	.45	.8930	.79	.6131
.12	.9928	.46	.8879	.80	.6000
.13	.9915	.47	.8827	.81	.5864
.14	.9902	.48	.8773	.82	.5724
.15	.9887	.49	.8717	.83	.5578
.16	.9871	.50	.8660	.84	.5426
.17	.9854	.51	.8617	.85	.5268
.18	.9837	.52	.8542	.86	.5103
.19	.9818	.53	.8480	.87	.4931
.20	.9798	.54	.8417	.88	.4750
.21	.9777	.55	.8352	.89	.4560
.22	.9755	.56	.8285	.90	.4359
.23	.9732	.57	.8216	.91	.4146
.24	.9708	.58	.8146	.92	.3919
.25	.9682	.59	.8074	.93	.3676
.26	.9656	.60	.8000	.94	.3412
.27	.9629	.61	.7924	.95	.3122
.28	.9600	.62	.7846	.96	.2800
.29	.9570	.63	.7766	.97	.2431
.30	.9539	.64	.7684	.98	.1990
.31	.9507	.65	.7599	.99	.1411
.32	.9474	.66	.7513	1.00	.0000
.33	.9440	.67	.7424		

TABLE OF SQUARES AND SQUARE ROOTS OF THE NUMBERS FROM 1 TO 100

Number	Square	Square Root	Number	Square	Square Root
1	1	1 000	51	26 01	7 141
2	4	1 414	52	27 04	7 211
3	9	1 732	53	28 09	7 280
4	16	2 000	54	29 16	7 348
5	25	2 236	55	30 25	7 416
6	36	2 449	56	31 36	7 483
7	49	2 646	57	32 49	7 550
8	64	2 828	58	33 64	7 616
9	81	3 000	59	34 81	7 681
10	1 00	3 162	60	36 00	7 746
11	1 21	3 317	61	37 21	7 810
12	1 44	3 464	62	38 44	7 874
13	1 69	3 606	63	39 69	7 937
14	1 96	3 742	64	40 96	8 000
15	2 25	3 873	65	42 25	8 062
16	2 56	4 000	66	43 56	8 124
17	2 89	4 123	67	44 89	8 185
18	3 24	4 243	68	46 24	8 246
19	3 61	4 359	69	47 61	8 307
20	4 00	4 472	70	49 00	8 367
21	4 41	4 583	71	50 41	8 426
22	4 84	4 690	72	51 84	8 485
23	5 29	4 796	73	53 29	8 544
24	5 76	4 899	74	54 76	8 602
25	6 25	5 000	75	56 25	8 660
26	6 76	5 099	76	57 76	8 718
27	7 29	5 196	77	59 29	8 775
28	7 84	5 292	78	60 84	8 832
29	8 41	5 385	79	62 41	8 888
30	9 00	5 477	80	64 00	8 944
31	9 61	5 568	81	65 61	9 000
32	10 24	5 657	82	67 24	9 055
33	10 89	5 745	83	68 89	9 110
34	11 56	5 831	84	70 56	9 165
35	12 25	5 916	85	72 25	9 220
36	12 96	6 000	86	73 96	9 274
37	13 69	6 083	87	75 69	9 327
38	14 44	6 164	88	77 44	9 381
39	15 21	6 245	89	79 21	9 434
40	16 00	6 325	90	81 00	9 487
41	16 81	6 403	91	82 81	9 539
42	17 64	6 481	92	84 64	9 592
43	18 49	6 557	93	86 49	9 644
44	19 36	6 633	94	88 36	9 695
45	20 25	6 708	95	90 25	9 747
46	21 16	6 782	96	92 16	9 798
47	22 09	6 856	97	94 09	9 849
48	23 04	6 928	98	96 04	9 899
49	24 01	7 000	99	98 01	9 950
50	25 00	7 071	100	1 00 00	10 000

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
101	1 02 01	10 050	151	2 28 01	12.288
102	1 04 04	10 100	152	2 31 04	12.329
103	1 06 09	10 149	153	2 34 09	12.369
104	1 08 16	10 198	154	2 37 16	12.410
105	1 10 25	10 247	155	2 40 25	12.450
106	1 12 36	10 296	156	2 43 36	12.490
107	1 14 49	10 344	157	2 46 49	12.530
108	1 16 64	10 392	158	2 49 64	12.570
109	1 18 81	10 440	159	2 52 81	12.610
110	1 21 00	10 488	160	2 56 00	12.649
111	1 23 21	10 536	161	2 59 21	12.689
112	1 25 44	10 583	162	2 62 44	12.728
113	1 27 69	10 630	163	2 65 69	12.767
114	1 29 96	10 677	164	2 68 96	12.806
115	1 32 25	10 724	165	2 72 25	12.845
116	1 34 56	10 770	166	2 75 56	12.884
117	1 36 89	10 817	167	2 78 89	12.923
118	1 39 24	10 863	168	2 82 24	12.961
119	1 41 61	10 909	169	2 85 61	13.000
120	1 44 00	10 954	170	2 89 00	13.038
121	1 46 41	11 000	171	2 92 41	13.077
122	1 48 84	11 045	172	2 95 84	13.115
123	1 51 29	11 091	173	2 99 29	13.153
124	1 53 76	11 136	174	3 02 76	13.191
125	1 56 25	11 180	175	3 06 25	13.229
126	1 58 76	11 225	176	3 09 76	13.266
127	1 61 29	11 269	177	3 13 29	13.304
128	1 63 84	11 314	178	3 16 84	13.342
129	1 66 41	11 358	179	3 20 41	13.379
130	1 69 00	11 402	180	3 24 00	13.416
131	1 71 61	11 446	181	3 27 61	13.454
132	1 74 24	11 489	182	3 31 24	13.491
133	1 76 89	11 533	183	3 34 89	13.528
134	1 79 56	11 576	184	3 38 56	13.565
135	1 82 25	11.619	185	3 42 25	13.601
136	1 84 96	11 662	186	3 45 96	13.638
137	1 87 69	11 705	187	3 49 69	13.675
138	1 90 44	11 747	188	3 53 44	13.711
139	1 93 21	11.790	189	3 57 21	13.748
140	1 96 00	11 832	190	3 61 00	13.784
141	1 98 81	11 874	191	3 64 81	13.820
142	2 01 64	11 916	192	3 68 64	13.856
143	2 04 49	11 958	193	3 72 49	13.892
144	2 07 36	12 000	194	3 76 36	13.928
145	2 10 25	12 042	195	3 80 25	13.964
146	2 13 16	12 083	196	3 84 16	14.000
147	2 16 09	12 124	197	3 88 09	14.036
148	2 19 04	12 166	198	3 92 04	14.071
149	2 22 01	12 207	199	3 96 01	14.107
150	2 25 00	12 247	200	4 00 00	14.142

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
201	4 04 01	14 177	251	6 30 01	15 843
202	4 08 04	14 213	252	6 35 04	15 875
203	4 12 09	14 248	253	6 40 09	15 906
204	4 16 16	14 283	254	6 45 16	15 937
205	4 20 25	14 318	255	6 50 25	15 969
206	4 24 36	14 353	256	6 55 36	16 000
207	4 28 49	14 387	257	6 60 49	16 031
208	4 32 64	14 422	258	6 65 64	16 062
209	4 36 81	14 457	259	6 70 81	16 093
210	4 41 00	14 491	260	6 76 00	16 125
211	4 45 21	14 526	261	6 81 21	16 155
212	4 49 44	14 560	262	6 86 44	16 186
213	4 53 69	14 595	263	6 91 69	16 217
214	4 57 96	14 629	264	6 96 96	16 248
215	4 62 25	14 663	265	7 02 25	16 279
216	4 66 56	14 697	266	7 07 56	16 310
217	4 70 89	14 731	267	7 12 89	16 340
218	4 75 24	14 765	268	7 18 24	16 371
219	4 79 61	14 799	269	7 23 61	16 401
220	4 84 00	14 832	270	7 29 00	16 432
221	4 88 41	14 866	271	7 34 41	16 462
222	4 92 84	14 900	272	7 39 84	16 492
223	4 97 29	14 933	273	7 45 29	16 523
224	5 01 76	14 967	274	7 50 76	16 553
225	5 06 25	15 000	275	7 56 25	16 583
226	5 10 76	15 033	276	7 61 76	16 613
227	5 15 29	15 067	277	7 67 29	16 643
228	5 19 84	15 100	278	7 72 84	16 673
229	5 24 41	15 133	279	7 78 41	16 703
230	5 29 00	15 166	280	7 84 00	16 733
231	5 33 61	15 199	281	7 89 61	16 763
232	5 38 24	15 232	282	7 95 24	16 793
233	5 42 89	15 264	283	8 00 89	16 823
234	5 47 56	15 297	284	8 06 56	16 852
235	5 52 25	15 330	285	8 12 25	16 882
236	5 56 96	15 362	286	8 17 96	16 912
237	5 61 69	15 395	287	8 23 69	16 941
238	5 66 44	15 427	288	8 29 44	16 971
239	5 71 21	15 460	289	8 35 21	17 000
240	5 76 00	15 492	290	8 41 00	17 029
241	5 80 81	15 524	291	8 46 81	17 059
242	5 85 64	15 556	292	8 52 64	17 088
243	5 90 49	15 588	293	8 58 49	17 117
244	5 95 36	15 620	294	8 64 36	17 146
245	6 00 25	15 652	295	8 70 25	17 176
246	6 05 16	15 684	296	8 76 16	17 205
247	6 10 09	15 716	297	8 82 09	17 234
248	6 15 04	15 748	298	8 88 04	17 263
249	6 20 01	15 780	299	8 94 01	17 292
250	6 25 00	15 811	300	9 00 00	17 321

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
301	9 06 01	17 349	351	12 32 01	18 735
302	9 12 04	17 378	352	12 39 04	18.762
303	9 18 09	17 407	353	12 46 09	18.788
304	9 24 16	17 436	354	12 53 16	18.815
305	9 30 25	17 464	355	12 60 25	18.841
306	9 36 36	17 493	356	12 67 36	18 868
307	9 42 49	17 521	357	12 74 49	18.894
308	9 48 64	17 550	358	12 81 64	18 921
309	9 54 81	17 578	359	12 88 81	18 947
310	9 61 00	17 607	360	12 96 00	18 974
311	9 67 21	17 635	361	13 03 21	19 000
312	9 73 44	17 664	362	13 10 44	19 026
313	9 79 69	17 692	363	13 17 69	19.053
314	9 85 96	17 720	364	13 24 96	19.079
315	9 92 25	17.748	365	13 32 25	19.105
316	9 98 56	17 776	366	13 39 56	19.131
317	10 04 89	17 804	367	13 46 89	19 157
318	10 11 24	17 833	368	13 54 24	19 183
319	10 17 61	17 861	369	13 61 61	19.209
320	10 24 00	17 889	370	13 69 00	19.235
321	10 30 41	17 916	371	13 76 41	19 261
322	10 36 84	17 944	372	13 83 84	19 287
323	10 43 29	17 972	373	13 91 29	19 313
324	10 49 76	18 000	374	13 98 76	19 339
325	10 56 25	18 028	375	14 06 25	19 363
326	10 62 76	18 055	376	14 13 76	19 391
327	10 69 29	18 083	377	14 21 29	19 416
328	10 75 84	18 111	378	14 28 84	19 442
329	10 82 41	18 138	379	14 36 41	19 468
330	10 89 00	18 166	380	14 44 00	19.494
331	10 95 61	18 193	381	14 51 61	19 519
332	11 02 24	18 221	382	14 59 24	19 545
333	11 08 89	18 248	383	14 66 89	19.570
334	11 15 56	18 276	384	14 74 56	19 596
335	11 22 25	18 303	385	14 82 25	19.621
336	11 28 96	18 330	386	14 89 96	19 647
337	11 35 69	18 358	387	14 97 69	19 672
338	11 42 44	18 385	388	15 05 44	19 698
339	11 49 21	18.412	389	15 13 21	19.723
340	11 56 00	18 439	390	15 21 00	19.748
341	11 62 81	18 466	391	15 28 81	19.774
342	11 69 64	18 493	392	15 36 64	19.799
343	11 76 49	18 520	393	15 44 49	19.824
344	11 83 36	18 547	394	15 52 36	19 849
345	11 90 25	18 574	395	15 60 25	19.875
346	11 97 16	18 601	396	15 68 16	19 900
347	12 04 09	18 628	397	15 76 09	19 925
348	12 11 04	18 655	398	15 84 04	19.950
349	12 18 01	18 682	399	15 92 01	19.975
350	12 25 00	18.708	400	16 00 00	20 000

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
401	16 03 01	20 025	451	20 34 01	21 237
402	16 16 04	20 050	452	20 43 04	21 260
403	16 24 09	20 075	453	20 52 09	21 284
404	16 32 16	20 100	454	20 61 16	21 307
405	16 40 25	20 125	455	20 70 25	21 331
406	16 48 36	20 149	456	20 79 36	21 354
407	16 56 49	20 174	457	20 88 49	21 378
408	16 64 64	20 199	458	20 97 64	21 401
409	16 72 81	20 224	459	21 06 81	21 424
410	16 81 00	20 248	460	21 16 00	21 448
411	16 89 21	20 273	461	21 25 21	21 471
412	16 97 44	20 298	462	21 34 44	21 494
413	17 05 69	20 322	463	21 43 69	21 517
414	17 13 96	20 347	464	21 52 96	21 541
415	17 22 25	20 372	465	21 62 25	21 564
416	17 30 56	20 396	466	21 71 56	21 587
417	17 38 89	20 421	467	21 80 89	21 610
418	17 47 24	20 445	468	21 90 24	21 633
419	17 55 61	20 469	469	21 99 61	21 656
420	17 64 00	20 494	470	22 09 00	21 679
421	17 72 41	20 518	471	22 18 41	21 703
422	17 80 84	20 543	472	22 27 84	21 726
423	17 89 29	20 567	473	22 37 29	21 749
424	17 97 76	20 591	474	22 46 76	21 772
425	18 06 25	20 616	475	22 56 25	21 794
426	18 14 76	20 640	476	22 65 76	21 817
427	18 23 29	20 664	477	22 75 29	21 840
428	18 31 84	20 688	478	22 84 84	21 863
429	18 40 41	20 712	479	22 94 41	21 886
430	18 49 00	20 736	480	23 04 00	21 909
431	18 57 61	20 761	481	23 13 61	21 932
432	18 66 24	20 785	482	23 23 24	21 954
433	18 74 89	20 809	483	23 32 89	21 977
434	18 83 56	20 833	484	23 42 56	22 000
435	18 92 25	20 857	485	23 52 25	22 023
436	19 00 96	20 881	486	23 61 96	22 045
437	19 09 69	20 905	487	23 71 69	22 068
438	19 18 44	20 928	488	23 81 44	22 091
439	19 27 21	20 952	489	23 91 21	22 113
440	19 36 00	20 976	490	24 01 00	22 136
441	19 44 81	21 000	491	24 10 81	22 159
442	19 53 64	21 024	492	24 20 64	22 181
443	19 62 49	21 048	493	24 30 49	22 204
444	19 71 36	21 071	494	24 40 36	22 226
445	19 80 25	21 095	495	24 50 25	22 249
446	19 89 16	21 119	496	24 60 16	22 271
447	19 98 09	21 142	497	24 70 09	22 293
448	20 07 04	21 166	498	24 80 04	22 316
449	20 16 01	21 190	499	24 90 01	22 338
450	20 25 00	21 213	500	25 00 00	22 361

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
501	25 10 01	22 383	551	30 36 01	23 473
502	25 20 04	22 405	552	30 47 04	23 495
503	25 30 09	22 428	553	30 58 09	23 516
504	25 40 16	22 450	554	30 69 16	23 537
505	25 50 25	22 472	555	30 80 25	23 558
506	25 60 36	22 494	556	30 91 36	23 580
507	25 70 49	22 517	557	31 02 49	23 601
508	25 80 64	22 539	558	31 13 64	23 622
509	25 90 81	22 561	559	31 24 81	23 643
510	26 01 00	22 583	560	31 36 00	23 664
511	26 11 21	22 605	561	31 47 21	23 685
512	26 21 44	22 627	562	31 58 44	23 707
513	26 31 69	22 650	563	31 69 69	23 728
514	26 41 96	22 672	564	31 80 96	23 749
515	26 52 25	22 694	565	31 92 25	23 770
516	26 62 56	22 716	566	32 03 56	23 791
517	26 72 89	22 738	567	32 14 89	23 812
518	26 83 24	22 760	568	32 26 24	23 833
519	26 93 61	22 782	569	32 37 61	23 854
520	27 04 00	22 804	570	32 49 00	23 875
521	27 14 41	22 825	571	32 60 41	23 896
522	27 24 84	22 847	572	32 71 84	23 917
523	27 35 29	22 869	573	32 83 29	23 937
524	27 45 76	22 891	574	32 94 76	23 958
525	27 56 25	22 913	575	33 06 25	23 979
526	27 66 76	22 935	576	33 17 76	24 000
527	27 77 29	22 956	577	33 29 29	24 021
528	27 87 84	22 978	578	33 40 84	24 042
529	27 98 41	23 000	579	33 52 41	24 062
530	28 09 00	23 022	580	33 64 00	24 083
531	28 19 61	23 043	581	33 75 61	24 104
532	28 30 24	23 065	582	33 87 24	24 125
533	28 40 89	23 087	583	33 98 89	24 145
534	28 51 56	23 108	584	34 10 56	24 166
535	28 62 25	23 130	585	34 22 25	24 187
536	28 72 96	23 152	586	34 33 96	24 207
537	28 83 69	23 173	587	34 45 69	24 228
538	28 94 44	23 195	588	34 57 44	24 249
539	29 05 21	23 216	589	34 69 21	24 269
540	29 16 00	23 238	590	34 81 00	24 290
541	29 26 81	23 259	591	34 92 81	24 310
542	29 37 64	23 281	592	35 04 64	24 331
543	29 48 49	23 302	593	35 16 49	24 352
544	29 59 36	23 324	594	35 28 36	24 372
545	29 70 25	23 345	595	35 40 25	24 393
546	29 81 16	23 367	596	35 52 16	24 413
547	29 92 09	23 388	597	35 64 09	24 434
548	30 03 04	23 409	598	35 76 04	24 454
549	30 14 01	23 431	599	35 88 01	24 474
550	30 25 00	23 452	600	36 00 00	24 495

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square F . . .
601	36 12 01	24 515	651	42 38 01	25 515
602	36 24 04	24 536	652	42 51 04	25 534
603	36 36 09	24 556	653	42 64 09	25 554
604	36 48 16	24 576	654	42 77 16	25 573
605	36 60 25	24 597	655	42 90 25	25 593
606	36 72 36	24 617	656	43 03 36	25 612
607	36 84 49	24 637	657	43 16 49	25 632
608	36 96 64	24 658	658	43 29 64	25 652
609	37 08 81	24 678	659	43 42 81	25 671
610	37 21 00	24 698	660	43 56 00	25 690
611	37 33 21	24 718	661	43 69 21	25 710
612	37 45 44	24 739	662	43 82 44	25 729
613	37 57 69	24 759	663	43 95 69	25 749
614	37 69 96	24 779	664	44 08 96	25 768
615	37 82 25	24 799	665	44 22 25	25 788
616	37 94 56	24 819	666	44 35 56	25 807
617	38 06 89	24 839	667	44 48 89	25 826
618	38 19 24	24 860	668	44 62 24	25 846
619	38 31 61	24 880	669	44 75 61	25 865
620	38 44 00	24 900	670	44 89 00	25 884
621	38 56 41	24 920	671	45 02 41	25 904
622	38 68 84	24 940	672	45 15 84	25 923
623	38 81 29	24 960	673	45 29 29	25 942
624	38 93 76	24 980	674	45 42 76	25 962
625	39 06 25	25 000	675	45 56 25	25 981
626	39 18 76	25 020	676	45 69 76	26 000
627	39 31 29	25 040	677	45 83 29	26 019
628	39 43 84	25 060	678	45 96 84	26 038
629	39 56 41	25 080	679	46 10 41	26 058
630	39 69 00	25 100	680	46 24 00	26 077
631	39 81 61	25 120	681	46 37 61	26 096
632	39 94 24	25 140	682	46 51 24	26 115
633	40 06 89	25 159	683	46 64 89	26 134
634	40 19 56	25 179	684	46 78 56	26 153
635	40 32 25	25 199	685	46 92 25	26 173
636	40 44 96	25 219	686	47 05 96	26 192
637	40 57 69	25 239	687	47 19 69	26 211
638	40 70 44	25 259	688	47 33 44	26 230
639	40 83 21	25 278	689	47 47 21	26 249
640	40 96 00	25 298	690	47 61 00	26 268
641	41 08 81	25 318	691	47 74 81	26 287
642	41 21 64	25 338	692	47 88 64	26 306
643	41 34 49	25 357	693	48 02 49	26 325
644	41 47 36	25 377	694	48 16 36	26 344
645	41 60 25	25 397	695	48 30 25	26 363
646	41 73 16	25 417	696	48 44 16	26 382
647	41 86 09	25 436	697	48 58 09	26 401
648	41 99 04	25 456	698	48 72 04	26 420
649	42 12 01	25 475	699	48 86 01	26 439
650	42 25 00	25 495	700	49 00 00	26 458

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
701	49 14 01	26 476	751	56 40 01	27 404
702	49 28 04	26 495	752	56 55 04	27 423
703	49 42 09	26 514	753	56 70 09	27 441
704	49 56 16	26 533	754	56 85 16	27 459
705	49 70 25	26 552	755	57 00 25	27 477
706	49 84 36	26 571	756	57 15 36	27 495
707	49 98 49	26 589	757	57 30 49	27 514
708	50 12 64	26 608	758	57 45 64	27 532
709	50 26 81	26 627	759	57 60 81	27 550
710	50 41 00	26 646	760	57 76 00	27 568
711	50 55 21	26 665	761	57 91 21	27 586
712	50 69 44	26 683	762	58 06 44	27 604
713	50 83 69	26 702	763	58 21 69	27 622
714	50 97 96	26 721	764	58 36 96	27 641
715	51 12 25	26 739	765	58 52 25	27 659
716	51 26 56	26 758	766	58 67 56	27 677
717	51 40 89	26 777	767	58 82 89	27 695
718	51 55 24	26 796	768	58 98 24	27 713
719	51 69 61	26 814	769	59 13 61	27 731
720	51 84 00	26 833	770	59 29 00	27 749
721	51 98 41	26 851	771	59 44 41	27 767
722	52 12 84	26 870	772	59 59 84	27 785
723	52 27 29	26 889	773	59 75 29	27 803
724	52 41 76	26 907	774	59 90 76	27 821
725	52 56 25	26 926	775	60 06 25	27 839
726	52 70 76	26 944	776	60 21 76	27 857
727	52 85 29	26 963	777	60 37 29	27 875
728	52 99 84	26 981	778	60 52 84	27 893
729	53 14 41	27 000	779	60 68 41	27 911
730	53 29 00	27 019	780	60 84 00	27 928
731	53 43 61	27 037	781	60 99 61	27 946
732	53 58 24	27 055	782	61 15 24	27 964
733	53 72 89	27 074	783	61 30 89	27 982
734	53 87 56	27 092	784	61 46 56	28 000
735	54 02 25	27 111	785	61 62 25	28 018
736	54 16 96	27 129	786	61 77 96	28 036
737	54 31 69	27 148	787	61 93 69	28 054
738	54 46 44	27 166	788	62 09 44	28 071
739	54 61 21	27 185	789	62 25 21	28 089
740	54 76 00	27 203	790	62 41 00	28 107
741	54 90 81	27 221	791	62 56 81	28 125
742	55 05 64	27 240	792	62 72 64	28 142
743	55 20 49	27 258	793	62 88 49	28 160
744	55 35 36	27 276	794	63 04 36	28 178
745	55 50 25	27 295	795	63 20 25	28 196
746	55 65 16	27 313	796	63 36 16	28 213
747	55 80 09	27 331	797	63 52 09	28 231
748	55 95 04	27 350	798	63 68 04	28 249
749	56 10 01	27 368	799	63 84 01	28 267
750	56 25 00	27 386	800	64 00 00	28 284

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
801	64 16 01	28 302	851	72 42 01	29 172
802	64 32 04	28 320	852	72 59 04	29 189
803	64 48 09	28 337	853	72 76 09	29 206
804	64 64 16	28 355	854	72 93 16	29 223
805	64 80 25	28 373	855	73 10 25	29 240
806	64 96 36	28 390	856	73 27 36	29 257
807	65 12 49	28 408	857	73 44 49	29 275
808	65 28 64	28 425	858	73 61 64	29 292
809	65 44 81	28 443	859	73 78 81	29 309
810	65 61 00	28 460	860	73 96 00	29 326
811	65 77 21	28 478	861	74 13 21	29 343
812	65 93 44	28 496	862	74 30 44	29 360
813	66 09 69	28 513	863	74 47 69	29 377
814	66 25 96	28 531	864	74 64 96	29 394
815	66 42 25	28 548	865	74 82 25	29 411
816	66 58 56	28 566	866	74 99 56	29 428
817	66 74 89	28 583	867	75 16 89	29 445
818	66 91 24	28 601	868	75 34 24	29 462
819	67 07 61	28 618	869	75 51 61	29 479
820	67 24 00	28 636	870	75 69 00	29 496
821	67 40 41	28 653	871	75 86 41	29 513
822	67 56 84	28 671	872	76 03 84	29 530
823	67 73 29	28 688	873	76 21 29	29 547
824	67 89 76	28 705	874	76 38 76	29 563
825	68 06 25	28 723	875	76 56 25	29 580
826	68 22 76	28 740	876	76 73 76	29 597
827	68 39 29	28 758	877	76 91 29	29 614
828	68 55 84	28 775	878	77 08 84	29 631
829	68 72 41	28 792	879	77 26 41	29 648
830	68 89 00	28 810	880	77 44 00	29 665
831	69 05 61	28 827	881	77 61 61	29 682
832	69 22 24	28 844	882	77 79 24	29 698
833	69 38 89	28 862	883	77 96 89	29 715
834	69 55 56	28 879	884	78 14 56	29 732
835	69 72 25	28 896	885	78 32 25	29 749
836	69 88 96	28 914	886	78 49 96	29 766
837	70 05 69	28 931	887	78 67 69	29 783
838	70 22 44	28 948	888	78 85 44	29 799
839	70 39 21	28 965	889	79 03 21	29 816
840	70 56 00	28 983	890	79 21 00	29 833
841	70 72 81	29 000	891	79 38 81	29 850
842	70 89 64	29 017	892	79 56 64	29 866
843	71 06 49	29 034	893	79 74 49	29 883
844	71 23 36	29 052	894	79 92 36	29 900
845	71 40 25	29 069	895	80 10 25	29 916
846	71 57 16	29 086	896	80 28 16	29 933
847	71 74 09	29 103	897	80 46 09	29 950
848	71 91 04	29 120	898	80 64 04	29 967
849	72 08 01	29 138	899	80 82 01	29 983
850	72 25 00	29 155	900	81-00 00	30 000

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
901	81 18 01	30 017	951	90 44 01	30 838
902	81 36 04	30 033	952	90 63 04	30 854
903	81 54 09	30 050	953	90 82 09	30 871
904	81 72 16	30 067	954	91 01 16	30 887
905	81 90 25	30 083	955	91 20 25	30.903
906	82 08 36	30 100	956	91 39 36	30 919
907	82 26 49	30 116	957	91 58 49	30 935
908	82 44 64	30 133	958	91 77 64	30 952
909	82 62 81	30 150	959	91 96 81	30.968
910	82 81 00	30 166	960	92 16 00	30 984
911	82 99 21	30.183	961	92 35 21	31 000
912	83 17 44	30 199	962	92 54 44	31 016
913	83 35 69	30 216	963	92 73 69	31.032
914	83 53 96	30 232	964	92 92 96	31 048
915	83 72 25	30 249	965	93 12 25	31.064
916	83 90 56	30 265	966	93 31 56	31 081
917	84 08 89	30 282	967	93 50 89	31 097
918	84 27 24	30 299	968	93 70 24	31.113
919	84 45 61	30 315	969	93 89 61	31 129
920	84 64 00	30 332	970	94 09 00	31 145
921	84 82 41	30 348	971	94 28 41	31.161
922	85 00 84	30 364	972	94 47 84	31 177
923	85 19 29	30 381	973	94 67 29	31.193
924	85 37 76	30 397	974	94 86 76	31 209
925	85 56 25	30 414	975	95 06 25	31 225
926	85 74 76	30 430	976	95 25 76	31 241
927	85 93 29	30 447	977	95 45 29	31 257
928	86 11 84	30 463	978	95 64 84	31 273
929	86 30 41	30 480	979	95 84 41	31 289
930	86 49 00	30 496	980	96 04 00	31 305
931	86 67 61	30 512	981	96 23 61	31 321
932	86 86 24	30 529	982	96 43 24	31 337
933	87 04 89	30 545	983	96 62 89	31 353
934	87 23 56	30 561	984	96 82 56	31 369
935	87 42 25	30 578	985	97 02 25	31.385
936	87 60 96	30 594	986	97 21 96	31 401
937	87 79 69	30 610	987	97 41 69	31.417
938	87 98 44	30 627	988	97 61 44	31.432
939	88 17 21	30 643	989	97 81 21	31.448
940	88 36 00	30 659	990	98 01 00	31.464
941	88 54 81	30 676	991	98 20 81	31 480
942	88 73 64	30 692	992	98 40 64	31.496
943	88 92 49	30 708	993	98 60 49	31 512
944	89 11 36	30 725	994	98 80 36	31 528
945	89 30 25	30 741	995	99 00 25	31.544
946	89 49 16	30 757	996	99 20 16	31 559
947	89 68 09	30 773	997	99 40 09	31 575
948	89 87 04	30 790	998	99 60 04	31 591
949	90 06 01	30 806	999	99 80 01	31 607
950	90 25 00	30 822	1000	100 00 00	31 623

INDEX

- Accuracy, standards of, in computation, 10-13
- Ackerson, Luton, 317
- Age-scale, in the combining of test scores, 185-187
- Anastasi, A., 283, 328, 458
- Array, in a correlation table, 260, 263
- Attenuation, correction of correlation coefficient for, 334-336; assumptions underlying, 336-338
- Average, definition of, 17, of correlation coefficients, 283-284
See Mean, Median, and Mode.
- Average deviation or *AD*, calculation of, 38-41; calculation of, by the Short Method (assumed mean), 44-47; from the median, 48-49; use of, 59
- Bagley, W. C., 459
- Bar diagram, 92-94
- Barlow's Tables, 465
- Barr, A. S., 341
- Beta coefficients, in partial and multiple correlation, 454
- Bias in sampling. *See* Sampling.
- Bingham, W. V., 325
- Binomial expansion, use in probability, 102-105; 106-108; graphic representation of, 104-105
- Bi-serial correlation, 366-371
- Blakeman, J., test for linearity of regression, 400-402
- Bravais, work in correlation, 257
- Brigham, C. C., 129, 211
- Brown, Wm., 315, 337, 348, 383, 464
- Burks, B. S., 356, 449, 456
- Buros, F. C., and Buros, O. K., 81
- Burt, Cyril, 445
- Carothers, F. E., 189
- Central tendency, measures of, 17; reliability of measures of, 200-208
- Chesire, L., 375
- Chi-square test, as a measure of goodness of fit, 119-124; as a measure of association, 377-387
- Clark, E. L., 318
- Classification of measures into a frequency distribution, 4-6
- Class-interval. *See* Step-interval.
- Clayton, B., 317
- Coefficient, of alienation, 344; of determination, in the interpretation of r , 355; of variation, or V , 51-55
- Coefficient of correlation, meaning of, 251-253; as a ratio, 255-257; represented graphically, 261-264; computation of, from assumed means, 265-271; computation of, from means, 271-277; reliability of, 280-283; effect upon, of range of talent in the group, 303-305; interpretations of, 342-356
- Coin tossing, in experiments upon the laws of chance, 101-104
- Column diagram. *See* Histogram.

- Comparable measures, 177-178; standard or z -scores, 178-180; equivalent scores in a common distribution, 180-183; percentile ranks, 183-185; median mental age method, 185-187; weighting scores according to their variability, 187-190; converting scores into equivalent ranks, 190-191
- Comparison, of obtained distributions with normal probability curve, 98-99, 124-126; of groups in terms of overlapping, 136-138. *See also* Chi-square, Skewness, and Kurtosis.
- Computation, rules for, 12-13
- Conrad, H. S., 346
- Contingency, coefficient of (C), 387, methods of computing C , 388-393, relation of C to chi-square, 387; comparison of C with r , 391
- Continuous series, 1, tabulation of measures in, 2-9
- Coordinate axes, 63-64; use in a correlation table, 295-297
- Correlation, linear, 251-253; positive, negative, and zero, 253-254; expressed as a ratio, 255-257; graphic representation of, 261-265; construction of correlation tables, 258-260; product-moment method in, 265-277; charts for use in, 270; difference formula in, 278-279, effect of errors of observation upon, 334-338; rank methods of computing, 359-366; spurious, 457-460. *See also* Partial correlation and Multiple correlation
- Correlation-ratio (η), in non-linear relationship, 393-396; computation of, 396-399, probable error of, 399, correction of, 400, comparison with r to determine linearity of regression, 400-402
- Criterion, value of, in determining the validity of tests, 324-329; prediction of, in multiple regression equation, 450-454
- Cumulative errors, effect upon multiple R , 456
- Cumulative frequencies, method of computing, 73
- Cumulative frequency graph, construction of, 72-74; smoothing of, 75
- Cureton, E. E., 270, 271, 337, 348, 423
- Data, continuous and discrete, 1-2
- Davenport, C. B., 465
- Deciles. *See* Percentiles.
- Derryberry, M., 433
- Deviation. *See* Quartile deviation, Average deviation, and Standard deviation
- Dice throwing, in experiments on the laws of chance, 105-106
- Difference, reliability of, between measures of central tendency, 210-223; between measures of variability, 223-226; between percentages, 228; between r 's, 281-283. *See* Standard error and Probable error.
- Discrete series, 2, Short Method applied to, 55-59
- Distribution, frequency. *See* Frequency distribution.
- Dunlap, J. W., 229, 270, 271, 329, 337, 348, 423, 465
- Dvorak, August, 270, 396

- Edgerton, H. A., 299
 Ekas, M. P., 465
 Equation, of a straight line, 295;
 plotting of regression lines in
 correlation diagram, 296-297
 Error, curve of, 107. *See also*
 Normal curve.
 Errors, of sampling, 242-246;
 constant, 245-246; variable,
 245. *See also* Probable and
 Standard errors.
 Ezekiel, M., 354, 423, 456, 465
 Fisher, R. A., 282, 284, 378, 379,
 402, 465
 Flemming, C. W., 444
 Foot-rule (Spearman's) in correla-
 tion, 363-365
 Franzen, R., 53, 433, 449
 Frequency distribution, construc-
 tion of, 4-9; graphical represen-
 tation of, 62-72
 Frequency polygon, construction
 of, 64-67; smoothing of, 67-69;
 comparison with histogram,
 71-72
 Galton, F., 257, 291
 Garrett, H. E., 187, 283, 328, 342,
 458
 Gates, A. I., 444, 448
 Graphic presentation, principles
 of, 62-64; of correlation coeffi-
 cient, 261-265. *See also* Fre-
 quency polygon, Histogram,
 Cumulative frequency graph,
 Percentile curve or Ogive, Line
 graph, Bar diagram.
 Greene, Harry A., 465
 Griffin, H. D., 433
 Grouping, in tabulating a fre-
 quency distribution, 4; assump-
 tions in, 6-9
 Guilford, J. P., 464
 Hart, H., 228
 Hartshorne, H., 227, 329
 Heilman, J. D., 356
 Heterogeneity, effect of, upon
 correlation, 303-305; upon the
 reliability of measures, 200-201,
 245-246
 Hillegas, M. B., 160
 Histogram, 69-70; comparison of,
 with frequency polygon, 71-72
 Holzinger, K. J., 291, 317, 396,
 464, 465
 Homogeneity, 32-33; effect of,
 upon variability, 243
 Hull, C. L., 171, 181, 182, 299,
 343, 346, 350, 444, 456
 Index of reliability, 319-320
 Interval. *See* Step-interval.
 Jackson, J. D., 317
 Jones, D. C., 108, 464
 Jones, H. E., 343
 Karsten, K. G., 89
 Kelley, T. L., 116, 208, 218, 224,
 229, 283, 305, 314, 320, 325,
 329, 345, 365, 375, 388, 464
 Kelly, E. L., 318
 Kurtosis, calculation of, 117-118;
 standard error of, 230-231
 Kurtz, A. K., 229, 465
 Lauer, A. R., 271
 Lickert, R., 161
 Lindquist, E. F., 221, 465
 Line graphs, 90-91
 Line of means, plotting of, 295-
 297
 Linearity of regression, tests for,
 400-402
 Martin, G. B., 346
 May, M. A., 227, 329, 412, 440

- McCall, W. A., 152
- Mean, arithmetic, calculation of, for ungrouped data, 17; for frequency distribution, 18, by Short Method (assumed mean), 26-29, when to use, 29, from combined distributions, 191-193, reliability of, 200-206
- Mean deviation, or *MD*. *See* Average deviation.
- Mean variation, or *MV*. *See* Average deviation.
- Median, calculation of, from ungrouped scores, 20; from frequency distribution, 21-22; in special cases, 23, when to use, 29-30; when data are discrete, 57-58, reliability of, 206-208
- Midpoint of step, how to find, 8; as a representative of all of the scores on the step, 9
- Midscore, 20
- Milne-Thomson, L. M., computation tables of, 465
- Miner, J. R., tables of, 465
- Mode, calculation of, 25-26; when to use, 30
- Monroe, W. S., 464
- Moore, H. L., 451, 456
- Morton, R. L., 60
- Moving average, use in smoothing a curve, 67
- Multiple coefficient of correlation, *R*, 411-412; computation of, in a three-variable problem, 419-420; formulas for, 430-432; "chance" *R*, 431; in a four-variable problem, 443; effect upon, of adding tests to the regression equation, 443-445; value of, in analysis, 448-449; limitations to use of, 455-457
- Musselman, J. R., 459
- Non-linear relationship, 393-402
- Normal probability curve, 98, illustrations of, 98-99, deduction from binomial expansion, 100-105, in psychological measurement, 106-108, equation of, 108; properties of, 113-114; comparison of obtained distribution with, 124-126; use in solution of a variety of problems, 132-143; in scaling test items, 143-157, in product scales, 157-161; in scaling judgments, 166-168; in transmutation of orders of merit into units of amount, 168-173
- Normality, divergence of frequency distributions from, 119; measurement of divergence from, 120-124
- Numbers, rounded, 10; exact and approximate, 11
- Odell, C. W., 464
- Ogive. *See* Percentile curve.
- Otis, A. S., 89, 270, 464
- Overlapping, in the measurement of groups, 136-138; of elements or factors in correlation, 348-354
- Partial correlation, 409-412; illustration of, in a three-variable problem, 412-420; notation in, 422; formulas for, 420-430; models of four- and five-variable problems, 433-437; illustration of, in a four-variable problem, 437-443; value of, in description and analysis, 445-448; limitations to the use of, 455-457
- Paterson, D. G., 183, 325, 455
- Pearl, Raymond, 351, 353, 464

- Pearson, Karl, 119, 124, 257, 291, 374, 375, 390, 400
- Percentage, standard error of, 226-227; standard error of the difference between, 228-229
- Percentile, construction of curve of, 82-85, uses of curve, 85-89; ranks, computation of, 79-82; graphic method of finding ranks, 84-85; scale, use of, in combining test scores, 183-185, scale, disadvantages of, 185
- Percentiles, calculation of, 76-78; graphic method of finding, 79, 84
- Peters, C. C., 305, 375, 377, 390, 398, 402, 461, 465
- Phillips, F. M., 446
- Pintner, R., 183, 299, 458
- Probable error, relation to Q , 37, relation to other measures of variability in the normal distribution, 114; of the mean, 205-206, of the median, 206; of Q , 209; of σ , 208-209; tables for finding the reliability of the difference in terms of, 214; of the correlation coefficient, 280-281; of the difference between means, 215-219; of the difference between medians, 222, of the difference between r 's, 281-283
- Probable error, of estimate, 301-303; in partial and multiple correlation, 419, 429-430; of an obtained score, 320-322
- Product-moment method of finding r , 265-279
- Quartile deviation (Q), calculation of, 34-38; when to use, 59; reliability of, 209
- Quartiles, Q_1 and Q_3 , computation of, 35-37
- Range, as a measure of variability, 34, when to use, 59; influence upon the coefficient of correlation, 303-305
- Rank-difference method of computing correlation, 360-363; when to use, 365
- Ranks, transmutation of, into units of amount, 168-173
- Reavis, George, 447
- Regression coefficient, 292; in partial and multiple correlation, 426-427
- Regression equations, 289-291; in deviation form, 291-297; in score form, 297-299; formulas for, in partial and multiple correlation, 426-428; value of, in prediction and control, 450-454; limitations to use of, 455-457
- Relative variability, coefficient of, 52. *See also* Coefficient of variation.
- Reliability, meaning of, 198-199; of the mean, 200-206; of the median, 206-208; of Q , 209; of σ , 208-209; of a percentage, 226; of differences, 210-229; problems involving, 231-241; sampling and reliability, 242-246; of test scores, 311-324; index of, 319-320; dependence of coefficient of, upon the size and variability of the group, 322-324
- Remmers, H. H., 318
- Rietz, H. L. et al., 464
- Rosenow, Curt, 431
- Ruch, G. M., 271, 313, 317, 329
- Rugg, H. O., 89, 464

- Saffir, M., 375
- Sampling, random, 198-200, 242-246; representative, 242-243, selection in, 245-246; reliability and, 198-230
- Scaling, of test items, 143-151; of total scores, 151-157; of answers to a questionnaire, 161-166; of judgments or ratings, 166-168. *See also* Age-scale, Percentile scale, *T*-scale.
- Scatter diagram, 258-261
- Schiller, B., 216
- Schneck, M. R., 187, 328, 342
- Score, in continuous and in discrete series, 2-3
- Semi-interquartile range, 35. *See* Quartile deviation.
- Shen, E., 208, 316, 329, 337
- Sheppard's correction for grouping error, 43
- Shock, N. W., 318
- Significant figures, 10
- Skewness, measurement of, 115-117; standard error of measures of, 229-230
- Sommerville, R. C., 15
- Sorenson, Herbert, 464
- Spearman, C., 315, 335, 363; foot-rule, 363-365
- Spearman-Brown prophecy formula, 315-319
- Spurious correlation, 457; arising from heterogeneity, 457-458; of indices, 458-459; of averages, 459; in composites, 459-460
- Standard deviation or σ , calculation of, 41-42, corrected for grouping, 43; calculation of, by Short Method, 49-50; calculation from raw scores, 51; when to use, 59; of the sum or difference of corresponding values in two series of test scores, 193-194; from combined distributions, 192-193; reliability of, 208-209; estimation of true value, 331-332
- Standard error, of a mean, 200-206, of a median, 206-208, of σ , 208-209; of Q , 209; of the difference between means, 210-221; of the difference between medians, 221-223; table for finding the reliability of a difference in terms of, 213; of sampling and of measurement, 332-333; of a percentage, 226-227
- Standard error of an obtained score, 320-322; in interpretation of r , 346-348
- Standard error, of a percentage, 226-227; of the difference between two percentages, 228-229
- Standard error, of estimate, 300-301; in partial and multiple correlation, 429-430
- Standard or z -scores, 178-180; as equivalent scores in a common distribution, 180-183
- Step-interval, size of, 4; methods of expressing, 6-7; midpoint of, 8; upper and lower limits of, 8-9
- Stoddard, G. D., 271, 465
- Symonds, P. M., 278, 433
- Tables of frequencies of normal curve, in terms of σ , 110; in terms of PE , 111; ordinates of the normal curve, 127
- Tabulation, of measures in a frequency distribution, 4-6; in a correlation table, 258-261

- Tetrachoric correlation, 371-377
 Thomson, G., 337, 348, 383, 458, 464
 Thorndike, E. L., 128, 129, 201, 337, 343
 Thurstone, L. L., 53, 86, 151, 158, 271, 328, 375, 464
 Thurstone, T. G., 328
 Toops, H. A., 444
 Trabue, M. R., 207, 222
 Transmutation of measures, 157-173; of judgments, 157-161, of orders of merit, 168-173
 True measures, 198-199; estimation of, 333-334, standard errors of, 333-334
 Tryon, R. C., 348
 T-scale, 151-156; advantages of, 156-157
 Universal percentile graphs, 83
 Validity, measurement of, in a test, 324; in terms of criteria, 324-325; effect upon, of lengthening a test, 325-327; indirect measures of, 327-329; relation of, to reliability, 329-331
 VanVoorhis, W. R., 305, 375, 377, 390, 398, 402, 461, 465
 Variability, meaning of, 33; measures of, 34-44; coefficient of relative variability, 51-55; reliability of measures of, 208-209. *See also* Average deviation, Quartile deviation, Range, Standard deviation.
 Variance, in the interpretation of r , 354-356
 Walker, H. M., 219, 257, 464
 Wechsler, D., 60
 Weighting scores in terms of the variability of the test, 187-190
 Whitley, Mary T., 327
 Wilks, S. S., 221
 Williams, J. H., 89
 Willoughby, R. R., 329
 Woodworth, R. S., 350
 Woody, Clifford, 147
 Wright, Sewall, 356
 Yerkes, R. M., 325
 Yoakum, C. S., 325
 Yule, G. U., 43, 105, 201, 202, 390, 398, 430, 458, 465